DISCOVERY AND REPLICATION OF PATHWAY-BASED TRANS-EXPRESSION QUANTITATIVE TRAIT LOCI

By

Laura Katherine Wiley

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2014

Nashville, Tennessee

Approved:

Professor William S. Bush

Professor Joshua C. Denny

Professor Josh F Peterson

ACKNOWLEDGEMENTS

I would like to give sincere thanks to my mentor on this project: Will Bush. This project started as a random side project and for better or for worse turned into a multi-year project filled with many learning opportunities. Will, thank you for putting up with all the "yet another SPIA disaster" emails and drop in meetings. Your mentorship taught me valuable lessons in study design, asking the right question, and perseverance. Additional thanks to my committee, Josh Peterson and Josh Denny for their invaluable guidance and support – most notably the reminder to always translate this bioinformatics project back to human health!

My acknowledgments would be incomplete without thanks to the Department of Biomedical Informatics faculty, staff and students. As the official genetics interloper, you all have welcomed me with open arms and greatly enriched my graduate school experience. Specific thanks go to Cindy Gadd and Mark Frisse for their continued mentorship and support. A special thanks to Rischelle Jenkins. She is the heart and soul of the DBMI training program and is always ready with a smile and helping hand.

To my friend and family, thanks for the scholarly (and not so scholarly) discussions. From maintaining sanity to advice on data visualization and presentation, you greatly contributed to this work. I could not have done this project without your support!

Finally, thanks to Vanderbilt University and the NIH for funding my training (T32-GM080178).

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

aRNA/ cRNA ..... antisense/complementary RNA

CDCV ..... common disease, common variant hypothesis

cDNA ..... complementary DNA

CEU..... CEPH/European descent individuals from the International HapMap Project

CHB ..... Han Chinese individuals from the International HapMap Project

ChIP-chip ..... chromatin immunoprecipitation & microarray analysis

ChIP-seq..... chromatin immunoprecipitation & sequencing

Dlx2..... distal-less homeobox 2

DNA ..... deoxyribonucleic acid

EBV..... Epstein-Barr virus

ENCODE ..... Encyclopedia of DNA Elements

eQTL ..... expression quantitative trait loci

ER ..... endoplasmic reticulum

EVI-1..... ecotropic virus integration site 1

FDR..... false discovery rate

GEO ..... Gene Expression Omnibus

GIH ..... Gujarati Indian individuals from the International HapMap Project

GO..... Gene Ontology

GSEA ..... gene-set enrichment analysis

GWAS..... genome-wide association study

HDAC2 ..... histone deacetylase 2

HIF1 ..... hypoxia inducible factor

HNF4..... hepatocyte nuclear factor 4

hnRNA ..... heterogeneous nuclear RNA

JPT ..... Japanese individuals from the International HapMap Project

kb..... kilobases

KEGG ..... Kyoto Encyclopedia of Genes and Genomes

LCL ..... lymphoblastoid cell line

LD ..... linkage disequilibrium

LWK ..... Luhya individuals from the International HapMap Project

MAF ..... minor allele frequency

miRNA..... microRNA

MKK ..... Maasai individuals from the International HapMap Project

mRNA ..... messenger RNA

MXL..... Mexican descent individuals in Los Angeles from the International HapMap Project

NF-1 ..... nuclear factor 1

OCT-1 ..... POU domain, class 2, transcription factor 1

ORA ..... over-representation analysis

P_G ..... probability of observing the combination of P_{NDE} and P_{PERT}

P_{GFDR} ..... false discovery rate corrected combined P_{NDE} and P_{PERT} probability

P_{NDE}..... probability of observing the number of differentially expressed genes

P_{PERT}..... probability of observing the pathway net accumulated perturbation value

RNA ..... ribonucleic acid

RNA-seq ..... RNA sequencing

RPEC..... retinal pigment epithelial cells

SNP ..... single nucleotide polymorphism

SOX2..... sex determining region Y- box 2

SPIA..... signaling pathway impact analysis

TFBS .... transcription factor binding site

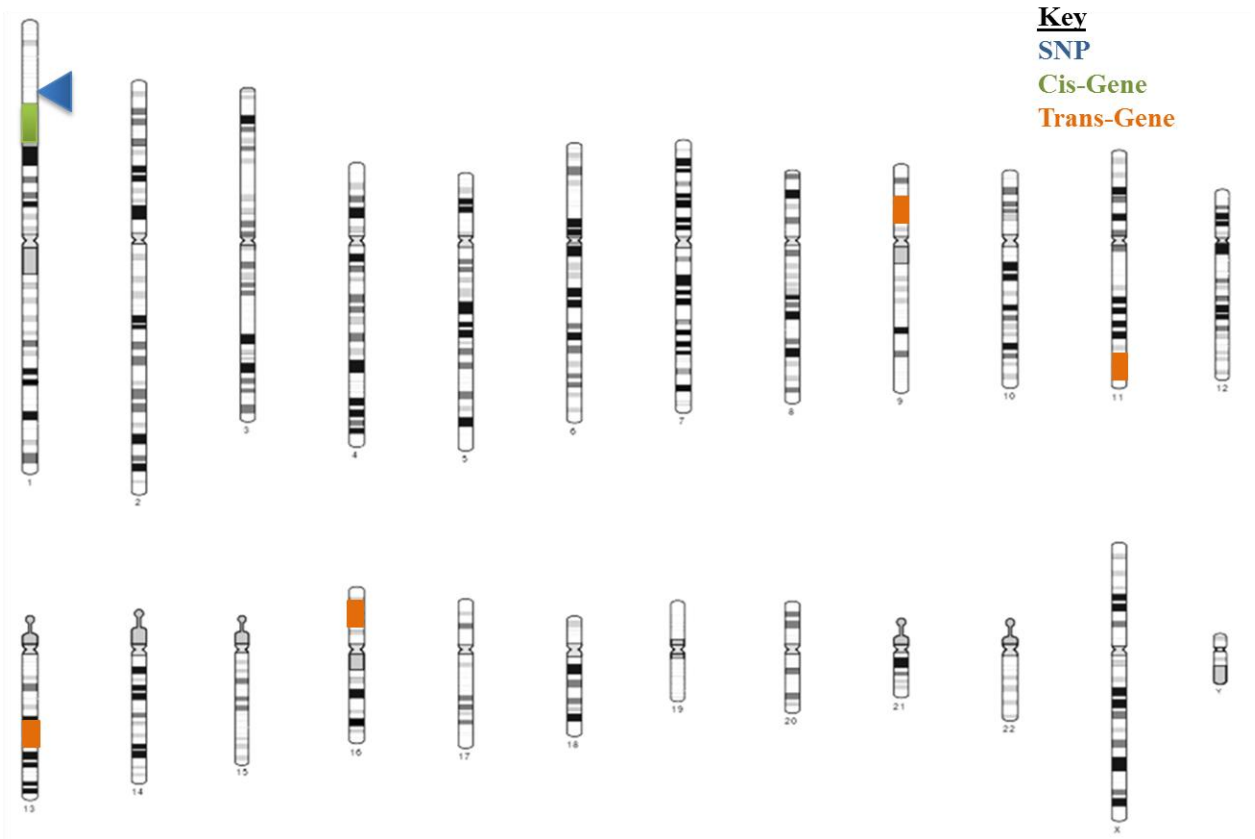
CHAPTER I

INTRODUCTION

Genome-wide association studies (GWAS) have identified numerous single nucleotide polymorphisms (SNP) associated with a variety of clinic traits [1]. Unfortunately, a full 88% of significant GWAS findings are in non-coding regions of the genome [2], making the exact biological mechanism underlying the association unclear. However, it has been shown that trait-associated SNPs are more likely to affect gene expression levels than non-trait associated SNPs [3] This effect – SNPs associating with differential gene expression levels – is termed expression quantitative trait loci or eQTL. There are two primary types of eQTL: cis and trans. Cis-eQTL are where the variant is within 500 kilobases (kb) of the gene that is differentially expressed. Trans-eQTL, on the other hand, are more than 500kb away from the differentially expressed gene, sometimes on entirely different chromosomes [4].

Most eQTL studies have focused on cis-eQTL for a variety of reasons. First, cis-eQTL have clear potential biological mechanisms. When a variant is located in a cis-regulatory region, it stands to reason that an allele change could alter the binding affinity of transcriptions factors resulting in differential expression. Other variants in non-coding portions of the gene could affect transcript stability again altering expression levels. Second, cis-eQTL are relatively easy to test. The relatively limited number of SNP-gene pairs allow for stringent, but not excessive, multiple testing penalties. Additionally, the number of tests is computationally tractable, especially compared to exhaustive testing for trans-eQTL which requires testing all possible combinations of SNPs and gene expression values. Further complicating matters, trans-eQTL have tended to have smaller effect sizes that are more variable across different tissue types than cis-eQTL [5]. Additionally trans-effects have been shown to replicate across populations less well than cis-eQTL, thereby requiring large sample sizes in populations of homogenous descent [6].Finally, trans-eQTL lack clear biological explanations for mechanisms of effect. Although some have proposed that close 3-dimensional contact between SNPs and the differentially expressed gene (as measured through chromatin conformation) could explain this effect [7], this hypothesis has not been extensively tested and the prevalence of this effect is unclear. We propose an alternate

A.



B.

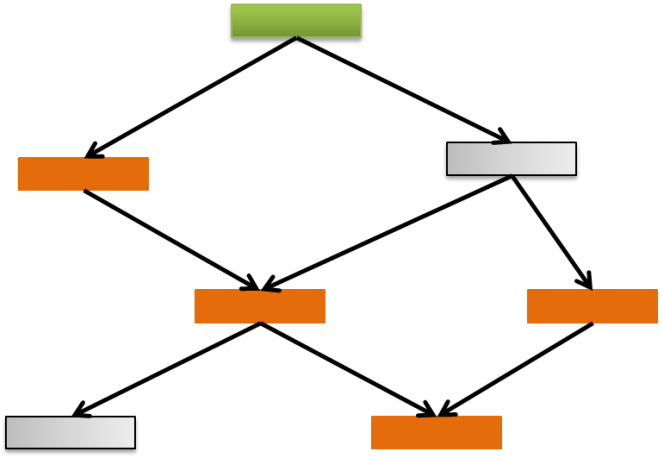


Figure 1. Study Rationale. Panel A shows all human chromosomes and illustrates a SNP (blue arrow) that has an effect on the expression of the cis-gene (green boxes), and also affects expression of four trans-genes on other chromosomes (orange boxes). Panel B demonstrates our studies hypothesis that these trans-genes are part of a biological pathway – explaining a potential mechanism for the SNP's trans-effects.

explanation that trans-eQTL may be operating through genetic pathways to alter gene expression.

Figure 1 explains our general conceptual framework for this problem. When there is a known cis-eQTL (represented by the blue arrow – SNP- and cis-gene – green rectangle), it stands to reason that this gene may be connected to trans-genes whose differential expression is associated to the SNP in question. If this is the case, it is likely that the SNP may be associated to the entire pathway – not just the genes inside the pathway. In this work we selected known cis-eQTL variants and tested them for association to differential expression of entire pathways. Given the general lack of replication and generalization of trans-effects, we performed both discovery and replication analyses in multiethnic cohorts. To further determine whether these effects act solely through the cis-eQTL gene (as is plausible in the hypothetical example) we also perform conditional analysis of each SNP-Pathway association, removing the effect of the expression of the cis-eQTL gene. Moreover we fully annotate all of our replicating SNPs with functional genomics data to further investigate potential effect mechanisms.

CHAPTER II

BACKGROUND

As described in the introduction, numerous single nucleotide polymorphisms have been associated with human traits. However to understand the mechanism behind those associations one needs to understand potential biological effects that can give rise to a disease state. Generically one can think of the following ways disease states can occur:

1. Improper protein formation or modification (affecting function or stability)
2. Improper trafficking or location of properly formed proteins
3. Improper expression of proteins (too much, too little, or inappropriate timing of expression).

To illustrate the first biological mechanism, SNPs in protein coding regions of the genome can alter the amino acid sequence of the protein perhaps affecting the proteins function or stability. Even intronic SNPs (i.e. non-coding variants) could be part of known splice site locations that alter exon arrangements again affecting the function or stability of the final protein product. For variants in unannotated regions of the genome, most have unknown function, but they likely contribute to the latter two potential molecular mechanisms underlying disease: improper trafficking or expression. In this work, we focus our attention on the third mechanism – altered gene expression. While some consider gene expression to mean the amount of fully functional protein product, in this work we refer to gene expression as the levels of RNA transcript in the cell. To better understand how single nucleotide polymorphisms can affect gene expression, we first need to understand how gene expression is measured and the basic biology behind normal gene expression.

Measuring Gene Expression Levels

RNA transcript abundance can be measured through a variety of experimental techniques. For high throughput, genome-wide assays, most experiments use microarray technologies or more recently RNA sequencing (RNA-seq). In both methods it is typical to amplify the transcript

abundance using either reverse transcription (making complementary DNA or cDNA) or antisense RNA amplification (making aRNA also known as complementary RNA or cRNA) [8]. Microarrays then hybridize these amplified sequences to labeled, preselected probes that are complementary to specific gene transcripts. Abundance of each transcript is measured base on the intensity of hybridization where higher intensity means more abundance of transcript. One disadvantage of this method is that it only captures transcripts included on the array. Additionally, variants located in the probe may artificially alter measured expression levels. In RNA sequencing, next generation sequencing technologies are used to sequence the amplified cDNA samples. Transcript abundance is usually inferred based on sequencing coverage, or read depth (i.e. how many copies of sequence map back to the original gene). Unlike microarrays, RNA-seq is agnostic in measuring each transcript, so the experiments are not limited to the probes used in the platform. Additionally, RNA-seq can identify specific transcript isoforms (differentially spliced transcripts – i.e. those with different numbers or ordering of exons).

Defining Functional Elements and their Role in Mediating Gene Expression

Transforming DNA into a protein product is a complicated process with strict regulation and numerous intermediate products. The expression of any single gene is dependent on cell type, temporal and biological conditions. For genes that are expressed, first the gene is transcribed into heterogeneous nuclear RNA (hnRNA) containing complete 3' and 5' untranslated regions, introns and exons. This hnRNA is processed into messenger RNA (mRNA) for transport outside of the nucleus by intron removal, appending a poly-adenine tail to the 3' end, and addition of a 5' methylguanosine cap. This mRNA can then be translated into a protein product by ribosomes outside of the nucleus. These protein products may be further modified into the final functional form of the protein. There are two clear biological routes to altered levels of RNA transcript abundance – 1) pre-transcription modifications affecting production of RNA products and 2) post-translational modifications that affect stability and therefore abundance of RNA transcript.

Pre-Transcriptional Regulation of Gene Expression

There are a number of factors affecting whether and how a gene will be transcribed. First, the gene and its regulatory elements must be accessible to transcription machinery. Second, alterations to regulatory elements or availability of specific regulatory elements can alter transcription levels. The openness of DNA can be accessed experimentally using DNaseI hypersensitivity assays [9]. DNaseI is an enzyme that cleaves DNA at pyrimidine bases in both single and double stranded DNA. However this enzyme can only act on open regions of DNA where the specific nucleotide can be interrogated. Thus identifying locations susceptible to DNaseI cleavage is a good proxy for open regions of chromatin.

Biological determination of DNA sequence availability to transcription machinery is primarily determined based on chromatin state. To have efficient packaging of DNA in a cell, DNA is wound around histone protein complexes to form a nucleosome. During replication, these nucleosomes can be further wound into fibers that are wound, compressed and coiled into chromatids. The tighter DNA is wound around the histone complexes, the less accessible the DNA sequence is for transcription. The levels of tightness of coiling around histones are determined by different types of modifications to the histone proteins. "Chromatin state" is therefore a qualitative descriptor of the types of histone modifications present and their effect on DNA accessibility.

Additionally, chromatin state can be indicative of the function of different regions of DNA. For instance, a study in 2010 used Hidden Markov Models to identify specific combinations of histone modifications that correlated with specific types of functions [10]. This analysis found that methylation of various lysine residues on histone 3 was associated with promoter elements – the site of RNA transcription initiation. Other histone combinations were correlated with different functions in intergenic regions, specifically: enhancers and insulators. Enhancers and insulators interact with transcription machinery to increase or decrease transcription respectively.

In regions that are open for transcription, different protein complexes or transcription factors bind to DNA to carry out transcription. Each transcription factor has slightly different regulatory processes dictating their effect. Thus knowledge of specific transcription factor binding sites can give insight into the patterns of expression of the gene. Typically transcription factors have a specific consensus sequence of nucleotide base pairs that correspond to the

physical connection between the transcription factor and the DNA sequence [11]. Sequence variation or nucleotide modifications in this region can alter the efficacy of transcription factor binding and therefore the amount of transcript produced.

The final type of functional variation that can impact the efficacy of regulatory machinery is DNA methylation. Unlike histone methylation that modifies the histone protein, DNA methylation involves addition of a methyl group to specific nucleotides. This methylation typically occurs in regions enriched for cytosine and guanine content and are termed CpG islands. When these nucleotides are methylated they tend to inhibit transcription of the nearby gene. These regions can also affect chromatin structure, again impacting relative abundance of transcription.

Post-Transcriptional Regulation of Gene Expression

During post-transcriptional modification, intron regions are removed and exons are spliced back together. However, for many genes the precise number or ordering of exons retained vary leading those genes to produce multiple different transcripts or isoforms. This alternative splicing is known to be highly heritable and common [12]. These splice isoforms are expressed at different levels and experience degradation at different frequencies due to nonsense mediated decay or other factors impacting transcript stability. These variations can have an artificial impact on total mRNA levels for the gene as measured through microarrays because some isoforms may bind more or less well to the expression probe. Another post-transcriptional modification, polyadenylation of the pre-mRNA, can affect transcript levels through reduced transcript stability.

Expression Quantitative Trait Loci

Cis-eQTL , or SNPs associated to the expression of a gene that is within 500-1000kb have been widely studied in humans [5,13-19]. Many of these studies have focused on characterizing the biological effects of these SNPs while others have focused on associations to diseases and human health [20-23]. It has even been shown that many of the disease associated variants in the GWAS catalog are known eQTL [3].

It is relatively clear how these variants could act to alter gene expression. As described by the overview of functional elements, most of the functional elements affecting gene expression are located close to the gene being expressed (i.e. within 500-1000kb up or downstream of the gene).Given that knowledge, it is likely that the cis-eQTL, or a variant in linkage disequilibrium (LD), could disrupt transcription factor binding, CpG methylation signals, splice site junctions or the polyadenylation signal sequence. What is less clear is how variants further away (or even on a different chromosome) could impact gene expression levels.

Trans-eQTL

Trans-eQTL are defined as variants associating with gene expression levels of a gene more than 1Mb away from the variant. Typically these effects are smaller and replicate less well than cis-variants [5,24]. In general these types of association are tested far less commonly, likely due to the lack of a clear biological mechanism an the dramatic expansion of statistical tests required to detect these effects [25]. However they do account for a significant portion of the heritability of gene expression levels and tend to replicate only in similar tissue types [26]. Investigation of known trait or disease associated SNPs found enrichment for trans-eQTL and cis-eQTL over non-associated common variants. Many variants associated to the same phenotypic trait all acted as trans-eQTL for the same trans-gene/s [27]. Trans-eQTL have also been directly associated to human disease [28].

Interestingly many trans-eQTL co-localize in the genome. These eQTL hotspots were initially identified in model organisms [29,30], and have recently been replicated in humans [31]. It is hypothesized that these variants may act as master regulators [32,33], though others propose that these variants are acting through pathway-based mechanisms [34-37]. In one study of the proposed pathway effect, known trans-eQTL for the same SNP were tested using Gene Set Enrichment Analysis (GSEA) and found an abundance of known upstream transcriptional regulators [34]. Other studies have used pathways and interaction networks to try to determine the actual genes being directly regulated by the SNP (i.e. removing those genes whose association is mediated through the pathway) [36,37]. One other study has proposed a similar approach to our work, performing pathway analysis over all known eQTL for a given SNP using Ingenuity Pathway Analysis (IPA) – a proprietary data source. This study was performed on a superset of European samples contained in the HapMap project and was not replicated in any

other sample or population [35]. Importantly, this study had numerous limitations. The lack of replication, generalization and the use of proprietary software/knowledge sources severely limit the impact of this work.

Controlling for Population Stratification in eQTL Studies

Generalizability (i.e. replication of effect across multi-ethnic populations) is important to determine the extensibility of the observed effect. However, when performing analyses in multiethnic populations, it is important to control for the confounding effect population can play. It is well understood that different distributions in minor allele frequency and phenotypic outcomes among different populations can decrease power and lead to spurious associations [38]. It has been shown that amongst the HapMap II populations, between 17% and 29% of genes are differentially expressed when comparing one population to another [6]. Given the known allelic differences and this phenotypic difference, it is critical to correct for population stratification in expression studies using multiple populations. Many studies perform all quality control and analysis steps separately in each population [6]. However this can greatly reduce power due to the smaller sample sizes within each population. Other strategies include correcting admixed populations gene expression levels with principal components analysis [15]. Still others have further corrected the gene expression levels within populations to align all populations to a standard (and hence comparable) distribution [14]. This approach allows all populations to be combined and analyzed at the same time, thereby increasing power through larger sample size. This correction involves performing a normal quantile transformation to each gene within a single population. While there is data loss in terms of the true spacing between relative gene expression values, the method reduces the effect of outlier expression values, and sets all populations and genes to the same distribution allowing for a combined analysis.

Approaches to Pathway Analysis

There are multiple types of pathway analyses used in bioinformatics. Many studies use ontologic approaches to group differentially expressed genes according to their Gene Ontology (GO) functions [39]. Others use biological pathways and protein interaction knowledge bases such as, Reactome [40,41] or the Kyoto Encyclopedia of Genes and Genomes (KEGG) [42,43].

Statistical analysis tends to use either over-representation analysis (ORA) or gene-set enrichment analysis (GSEA) Over-representation analysis essentially tests the hypothesis that there are more differentially expressed genes in a pathway or ontology group than expected by chance alone (typically determined through permutation testing [44]. Gene-set enrichment analysis is a more complicated approach that ranks genes from a given set (biological pathway, physical proximity, GO category, etc) based on the correlation of their expression with a particular phenotype. Under the null hypothesis that the pathway is unrelated to the phenotypic outcome, correlation values would be randomly distributed. GSEA uses a random walk algorithm to calculate an enrichment score that measures how overrepresented the set of genes are at either the top or bottom of the distribution. The significance of this enrichment score is determined through phenotypic permutation testing [45].

What both ORA and GSEA fail to take into account is the organization of the gene sets/pathways and topology features of those networks. Essentially all genes are treated with equal weight where in true biological pathways some genes have more impact than others (e.g. may impact expression of many genes). Similarly, differential expression of subsets of genes may provide more evidence of dysregulation than others. For instance if a set of interacting genes are all differentially expressed this is more relevant biologically than if the differentially expressed genes are randomly distributed throughout the pathway. It is for these reasons that Signaling Pathway Impact Analysis (SPIA) was developed [46].

Project Specific Data Sets and Tools

International HapMap Project and the 1000 Genomes Project

One hypothesis of genetic influences on complex traits is the common disease common variant hypothesis (CDCV). Essentially, if a disease is common within a population, it stands to reason that the genetic factors behind that disease would also common. In 2001 it was estimated that 10 million bases in the genome had common variation (alternate alleles at >1% frequency in the population) across the world human population [47]. While it was possible to interrogate some of this variation in regions of known interest, identifying candidate regions in an agnostic manner was not feasible due to sequencing cost and the burden of correction for the high number of statistical tests needed for single variant association. However, by understanding the process

by which genetic variation arises, it was determined that a much smaller subset of variants could be used to capture the broader common variation in the genome.

Variation in the genome typically originates in singular mutation events on a common genetic background. Over generations this mutation can propagate through a population and become common. This SNP is associated with the other variants from the background sequence. This region, called a haplotype, is changed in sequence and length as further mutation or recombination events occur. Linkage disequilibrium (LD) describes this phenomenon of co-inheritance of these alleles. Studies can then use variants that most precisely define the haplotype present (thereby tagging other variation in the region) for more efficient interrogation of the human genome. This approach has been widely used in genome-wide association studies. However, the amounts of LD present and even which SNPs tag the most variation differs by population. In 2003 the International HapMap Consortium was formed to create a human haplotype map for multiple ancestral populations [48].

The first phase of this project developed haplotype maps for four ancestral populations – specifically 30 trios (parents and one offspring, 90 individuals total) from the Yoruba in Ibadan, Nigeria (YRI), 30 trios from Utah (CEU), 45 unrelated Han Chinese from Beijing (CHB) and 45 unrelated Japanese in Tokyo (JPT). The first phase genotyped common variation (minor allele frequency (MAF) > 5% in the population) every 5 kilobases (kb) [49]. The second phase of this project increased the density of variation to one SNP every kilobase, capturing approximately 25-35% of common SNPs in the human genome [50]. The third phase of the project greatly increased the number of individuals and populations (1,184 individuals across 11 populations). Importantly, this project phase only captured approximately 1 million variants through genotyping technologies compared with the combined Phase I and II sequencing and genotyping efforts which measured over 3 million SNPs. The original four populations were included in the project (with new samples from these populations). Additional populations included: ASW – African Americans from the southwest United States; CHD – Chinese individuals in Denver, Colorado; GIH – Gujarati Indians from Houston, Texas; LWK – Luhya in Webuye, Kenya; MKK – Maasai in Kinyawa, Kenya; MXL – Mexican ancestry individuals from Los Angeles, California; TSI – Tuscans in Italy [51].

Ultimately the success of the International HapMap Project and the GWAS it enabled, led the field to begin examination of low frequency and rare variation (minor allele found in <1% of

the population) through the 1000 Genomes Project. This project sought to identify haplotype information for all genetic variation (not just common variation as in HapMap) for variants with a minor allele frequency > 1%. Additionally the project catalogued rare variation (minor allele frequency down to 0.1%) in coding regions of the genome. Like Phase III of the HapMap Project, this project studied a larger set of populations pulled from five broad ancestral groups. Ultimately the project concluded that it had identified over 95% of all currently accessible variation in an individual's genome [52]. In addition to the genetic data that these two projects produced, many of the individuals from the population had lymphoblastoid cell lines (LCLs) created through transformation of B-lymphocytes with Epstein Barr Virus (EBV). These cells are available for research and have been widely used in other large scale projects.

Functional Variation Databases

Following the completion of the Human Genome Project, it became obvious that identifying the function of the 3 billion bases was the next step forward for the field. This need created the Encyclopedia of DNA Elements Project (ENCODE) whose goal was to identify all functional elements in the human genome sequence [53]. In 2004, the pilot phase of this project started and focused on only 1% of the genome. The pilot phase sought to identify procedures and technologies that could be used to eventually interrogate the entire genome – developing new technologies where necessary. Ultimately the ENCODE consortium planned to identify the following types of functional elements in the genome:

- Genes
- Exons
- Origins of Replication
- Replication Termination Locations
- Transcription Factor Binding Sites
- Conserved Regions Across Species
- Chromatin Modifications
- Sites of Methylation
- DNaseI Hypersensitive Sites
- Promoters
- Enhancers
- Repressors/Silencers

These features would be measured using a combination of transcript and chromatin immunoprecipitation microarray hybridization (ChIP-chip), other array based technologies (for methylation), and computational methods.

The pilot project of ENCODE was completed and published in 2007. These preliminary results showed that the genome is pervasively transcribed and revealed numerous details about regulatory elements and sequences affecting gene transcription [54]. With the knowledge and tools from the pilot project in hand, the project extended their analysis to the remainder of the genome. This vastly expanded analysis culminated in a coordinated release of 30 different publications highlighted in a special issue of Nature [55]. In total the final ENCODE project evaluated up to seven major types of functional variation (DNA methylation, open chromatin regions, RNA binding sites, RNA transcript sequences, ChIP-seq, histone modifications, and transcription factor binding sites) in more than 150 cell lines. Not all types of annotation are available for every cell line, so there is more work to be completed.

CHAPTER III

METHODS

Datasets

Gene Expression

Given that gene expression offers an important intermediate link between genetic variation and disease and is highly heritable, Barbara Stranger and others used the genetic variation present in HapMap and the accompanying cell lines to perform studies of eQTL in these populations [6]. In 2007, this group measured gene expression of lymphoblastoid cell lines from 270 Phase I and II HapMap samples using the Illumina whole-genome expression (WG-6 version 1) array. This platform measured 47,294 probes, and after filtering probes that map to multiple positions in the genome, a total of 14,456 probes representing 13,643 genes remained. These data were made publically available through the Gene Expression Omnibus (GEO; Series Accession Number GSE6536; ref. 19). These data were further processed by Veyrieras et. al. 2008 to remove known probe errors (i.e. a SNP was located in the probe leading to spurious artifacts). We downloaded these further processed gene expression data from (http://eqtnminer.sourceforge.net/).

Following the successful completion of Phase III of the HapMap project, Stranger and colleagues extended their previous study to a more rigorous examination of the influence of ancestry on gene expression [15]. In this study, gene expression for lymphoblastoid cell lines from 726 individuals in the HapMap III project were measured using Illumina Sentrix Human-6 Expression BeadChip version 2. The populations measured included CEU, CHG, GIH, JPT, LWK, MEX, MKK, and YRI. In the populations overlapping the HapMap II samples, this study included a mix of samples from HapMap II and the new individuals added in HapMap III. In total, this group measured 47,294 probes, and the data were made freely available at Array Express (Series Accession Number E-MTAB-264).

Genotypes

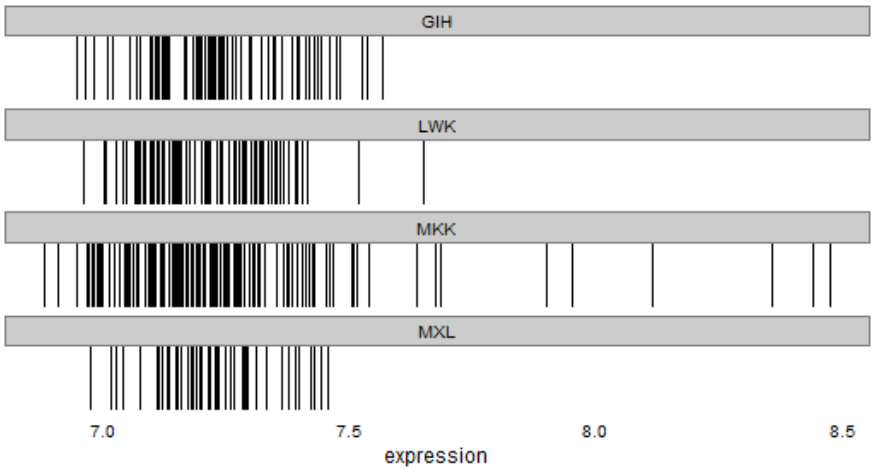
For the 207 individuals from HapMap Phase I and II, genotype data was download from release 24 of the International HapMap project. As described in Chapter 2, the third phase of the HapMap project did not genotype as many variants as the previous two phases. When comparing the SNPs available for the 207 HapMap I and II samples to those available for the 466 Phase III HapMap samples, only about ~50% of these had been genotyped. We extract all available genotypes from draft release 2 of the International HapMap project for all 466 individuals However, 236 of those 466 individuals had also been sequenced as part of the 1000 genomes project. Groups have previously performed haplotype phasing (1000G Phase I version 3 MACH panels) on those data and have made those data freely available [49,56]. We drew the remaining half of the SNPs available in Phase I and II HapMap samples but not Phase III HapMap from this resource for the 236 individuals sequenced.

Study Populations

The discovery population consisted of 210 independent multiethnic samples, specifically 60 CEPH and 60 Yoruba parental samples, 45 Han Chinese and 45 Japanese unrelated individuals from the Phase II HapMap Project. Gene expression measures were from Stranger et.al. 2007 and genotype data came from release 24 of the International HapMap project.

The replication set consisted of up to 466 independent multiethnic samples, specifically 34 Han Chinese, 39 Japanese, 82 Gujarati Indians, 83 Luhya, 134 Maasai, 53 Yoruba and 42 individuals of Mexican descent in Los Angeles. These were all unrelated individuals from the Phase III HapMap Project and although some populations overlapped with our discovery set (CHB, JPT & YRI) the individuals tested were independent. Gene expression values were from Stranger et. al. 2012, 65 SNPs were draft release 2 of the International HapMap project for all 466 individuals, and 62 SNPs from 1000G Phase I version 3 MACH panels. Unfortunately, the SNPs from the 1000 genomes project only included CHB, JPT, LWK, MXL, and YRI individuals giving a total sample size of 236 individuals for those SNPs.

A.



B.

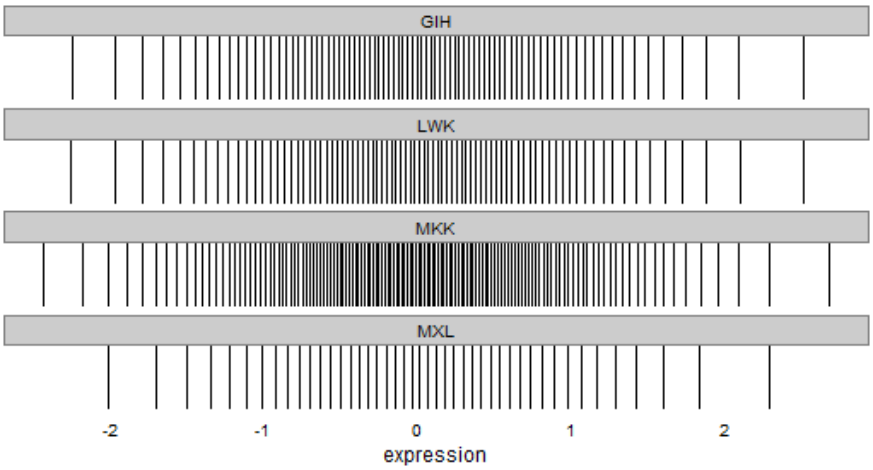


Figure 2. Normal quantile transformation of a single gene across HapMap III samples. Lines correspond to the gene expression value for a single individual and gene. Panel A shows the original raw gene expression values for each population while Panel B is following normal quantile transformation.

Normalization of Gene Expression Values

As part of normal quality control procedures, raw expression levels were normalized with quantile normalization within replicates and then median normalized across all samples. These quality control methods were performed by Stranger et. al. (2007 and 2012) prior to data downloads. However, to be able to combine populations with different distributions of gene

expression values, we had to apply additional normalization procedures. We used the normal quantile transformation originally proposed by Veyrieras et. al. In this method a single gene's expression value for each individual in a single population is ranked in numerical order. These rankings are then transformed by the following equation:

$$\frac{(r - 0.5)}{n}$$

where r is the rank of the gene expression value and n is the number of individuals in the population. The value produced by the equation corresponds to the quantile of the standard normal distribution that will be assigned as the gene expression value for that gene for that individual. Put more simply, this creates a ranking system where the transformed gene expression value depends on the relative expression value within the population and the size of the population tested. For an example of this transformation in HapMap III only populations, see **Figure 2**

Selection of Single Nucleotide Polymorphisms for Testing

We only chose to investigate SNPs that had a known cis-eQTL effect in lymphoblastoid cells lines (as measured in [14]). The original study defined all SNPs within 500kb upstream of the transcription start site and 500 kb downstream of the transcription end site as cis-variants. As shown in **Figure 3**, for each of the 744 genes that had at least one significant eQTL (as defined by a p-value < $7x10^{-6}$ which corresponds to a gene-level false discovery rate of 5%) we calculated a gene-specific significance threshold using a Bonferroni correction for the number of cis-SNPs. Cis-SNPs that met these significance thresholds were further filtered to identify the most significant SNP identifying independent loci for each gene. This filtering process relies on empirical estimates of linkage disequilibrium as calculated using the clump function in PLINK[57]. This algorithm requires four parameters – the significance cutoffs for index and clumped SNPs, an LD threshold and a physical distance threshold. We did not set a significance threshold for this procedure as all SNPs already met our significance cutoffs. We clumped SNPs that were within 250kb of the index SNP that also had an R^2 of at least 0.5. The most significant SNP from each clump was carried forward in the analysis. Any SNP that fell outside of a clump were also carried forward for analysis.

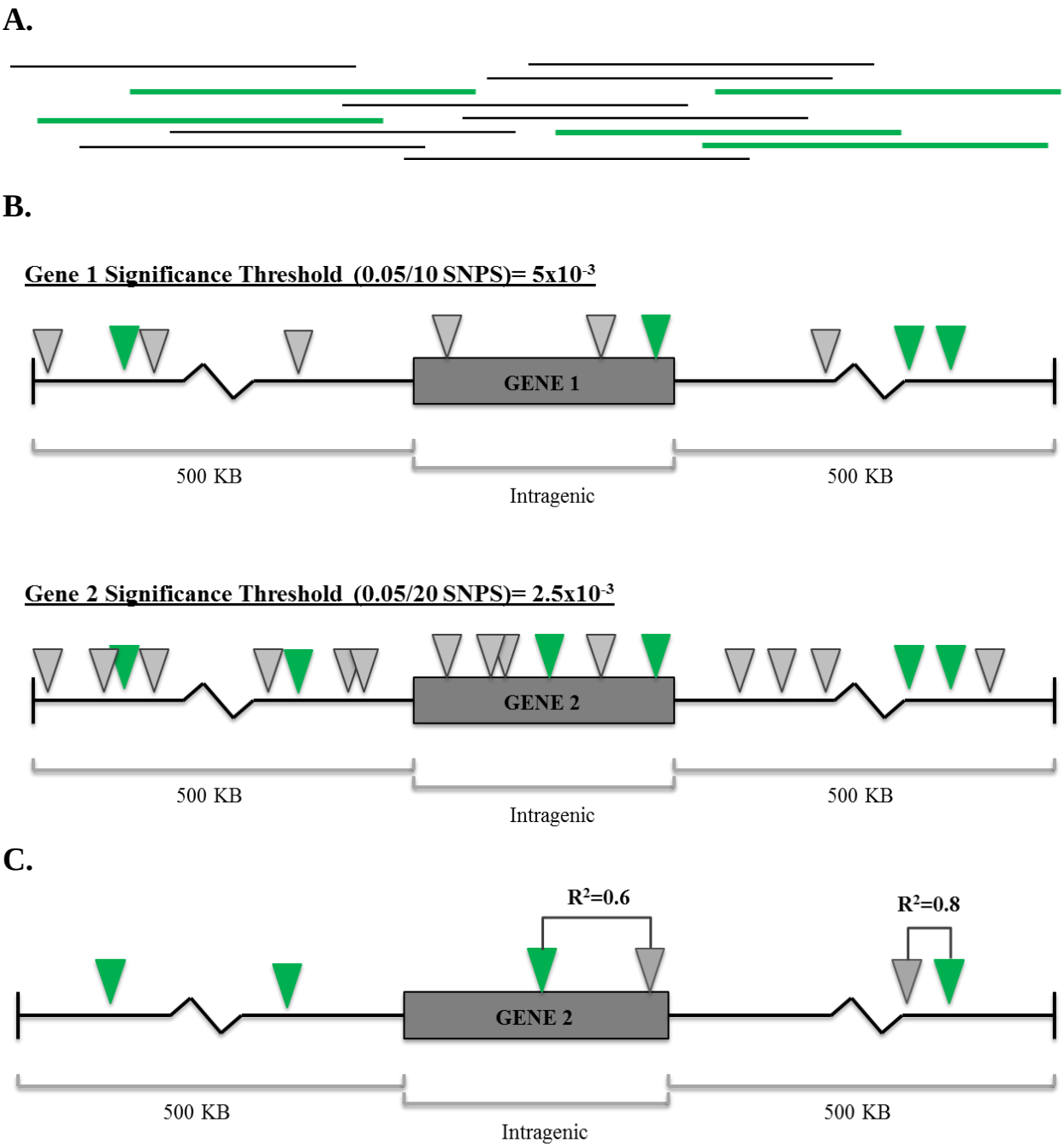


Figure 3. SNP Selection Procedure. Panel A illustrates that for all genes tested for cis-eQTL by Veyrieras *et. al.* we only selected those that had at least one eQTL meeting a gene-level false discovery rate of 5% (in green). Panel B summarizes the gene specific Bonferroni correction based on how many SNPs were tested for each gene independently. In Gene 2 there were more SNPs tested so the significance threshold was proportionally lowered. All SNPs passing this threshold (in green) were considered for the analysis. Finally in Panel C, for all significant SNPs in Gene 2 a test of linkage disequilibrium removed variants with a correlation higher than 0.5. In cases of LD the most significantly associate SNP was retained (colored green). SNPs not in LD with other variants were also retained (colored green).

Pathway-Based trans-eQTL Analysis

Overview of Signaling Pathway Impact Analysis

Signaling Pathway Impact Analysis calculates two probability measurements. The first is analogous to ORA, in that it calculates the probability of observing the number of differentially expressed genes in the pathway compared to the null hypothesis of random differential expression. This measure will be referred to as P_{NDE}. The second probability measure relates to perturbation of the pathway, essentially taking into account *which* genes in the pathway are differentially expressed. This probability is calculated with respect to a perturbation factor defined as:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)}$$

Here $\Delta E(g_i)$ is the normalized gene expression log fold change. This value is adjusted by the sum of the perturbation factor for each of the upstream genes $[PF(g_j)]$ after normalization for the number of downstream genes $[N_{ds}(g_j)]$. These summed perturbation factors are weighted based on the type of interaction with the current gene (e.g. activation, inhibition, etc). This weighting is captured using the β_{ij} term. Performing this calculation for each differentially expressed gene creates a large set of simultaneous equations that can be solved to calculate the perturbation factor value for each gene.

However, to capture the total true pathway perturbation (i.e. the accumulation of effect through the pathway) one cannot use the perturbation factor value alone. This is because the perturbation factors include the level of gene expression for each gene in the pathway. Instead we want to calculated perturbation accumulation – essentially the excess effects on expression that is propagated through interconnected differentially expressed genes. This measure is caclculated as:

$$Acc(g_i) = PF(g_i) - \Delta E(g_i)$$

This measure is then summed across the pathway creating a total accumulated perturbation measure. Using bootstrapping, the probability of observing a total accumulation as extreme or more extreme by chance alone is calculated and referred to as P_{PERT}. An visual example of these effects are presented in **Figure 4**.

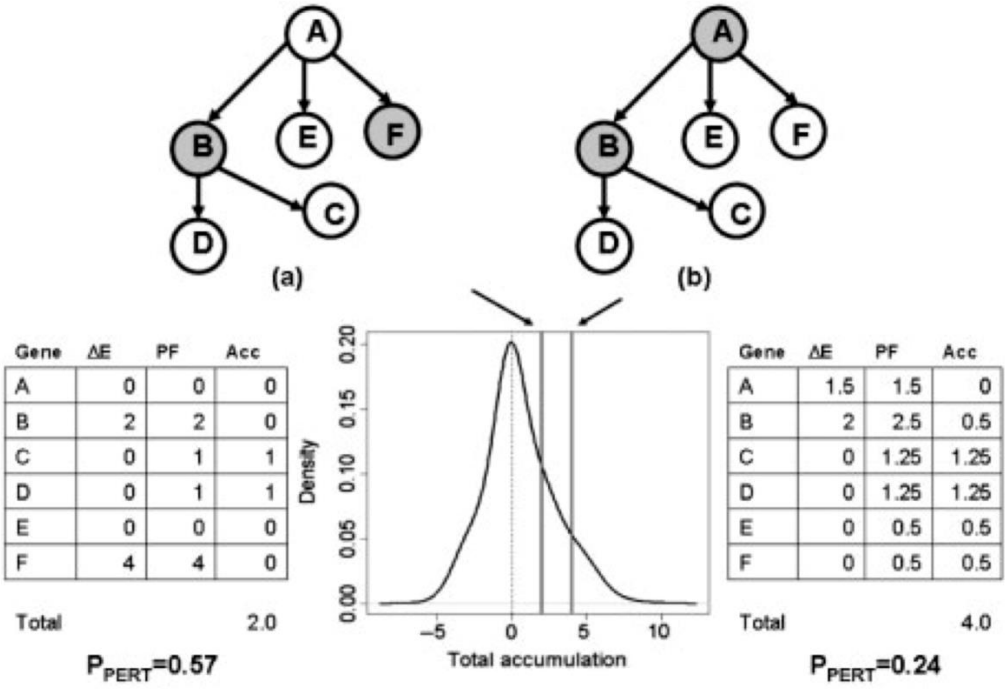


Figure 4. Example of SPIA Perturbation Measure. In this figure two different types of pathway dysregulation are compared. Both pathway (a) and (b) would have the same P_{NDE} values as they both have two of seven differentially expressed genes. However the perturbation statistic, P_{PERT} will be different. In pathway (a) there is a smaller total pathway expression accumulation because the two differentially expressed genes (grey) are not directly connected in the pathway. Although neither pathway has significant perturbation factors, pathway (b) has a smaller statistic because of the cumulative effects of the differentially expressed genes, Gene A and Gene B having direct contact in the pathway. Figure from [46].

Finally, the two measures, (P_{NDE} and P_{PERT}) are combined into a total probability score, P_G. For each pathway, i, the total probability score is calculated with reference to c_i where c_i is calculated as the product of P_{NDE} and P_{PERT} [e.g. $c_i = P_{NDE}(i) \cdot P_{PERT}(i)$]. The calculate for P_G is as follows:

$$P_G = c_i - c_i \cdot \ln(c_i)$$

In all cases, the probability measures are independent of pathway size, which allows for direct comparison of values among pathways.

SPIA can be applied to any pathway database that has interconnected nodes connected through directed edges. The original implementation of this approach uses human KEGG pathways which include a number of cellular and disease pathways. In total there are more than

60 human pathways in KEGG, which means that it is necessary to correct probability measures calculated in SPIA for multiple testing. Given that these pathways are highly correlated, controlling for a 5% false discovery rate (FDR) [58] is appropriate and implemented in the original SPIA R package. This measure will be referred to as P_{GFDR}.

Approach

For each SNP that passed our filtering process, we analyzed the effect of that SNP on differential expression of KEGG pathways using the tool Signaling Pathway Impact Analysis (SPIA). [46] To determine the effect of each SNP on the various pathways we first regressed gene expression values onto the additively encoded genotype for each SNP independently using linear regression. Genes with a regression p-value <0.05 were considered differentially expressed. Because of the normalization procedures used to combine our populations, we could not use the typical log-fold change in gene expression for this analysis. Instead, we used the raw beta value from the SNP regression – essentially representing the per-allele additive effect on gene expression.

Significance Thresholds

As mentioned previously, SPIA examines all KEGG pathways an FDR corrected p-value (P_{GFDR}) is presented to account for multiple testing at a pathway level. However, because we are performing multiple pathway analyses – one for each SNP tested – we need to perform additional correction. Each SNP is independent (due to the linkage disequilibrium filtering), so we corrected the P_{GFDR} value using a Bonferroni correction for each of the 2909 SNPs tested resulting in a threshold of $p < 1.7x10^{-5}$. All SNP-Pathway associations with a P_{GFDR} below this threshold were tested for replication in the replication population. Given that we are only testing certain SNP-pathway combinations, we will use the unadjusted global p-value (P_G) in our replication thresholds. We considered SNP-Pathway combinations with nominally significant ($p<0.05$) P_G values as replicated. However, to identify SNP-pathway effects that appear to be especially robust, we calculated a Bonferroni corrected threshold for the number of SNP-pathway combinations tested (n=221) a P_G threshold of $2.3x10^{-4}$.

Identification of Possible Mechanisms of Action

Investigating Replication of Known cis-eQTL

Given that we selected SNPs on the basis of their purported cis-eQTL effects found by Veyrieras *et. al*. in our discovery population, we were interested to see if these primary effects were able to be replicated. For each of the SNPs with significant SNP-Pathway replication, we tested the effect of the SNP on the cis-gene's expression. Using linear regression, we associated additively encoded SNPs on the gene expression value. For SNPs that had multiple cis-eQTL genes in our discovery population, we performed regression for the SNP on each gene. Given that this is a replication, we used a liberal, nominally significant threshold ($p<0.05$).

Removing Effect of cis-eQTL Gene Expression

To determine whether these pathway-based expression changes are propagated through the cis-eQTL gene alone or if there are extra trans-effects by these SNPs, we analyzed all SNP-Pathway combinations that replicated across both datasets. As with the discovery and replication analyses, we regressed the gene expression values onto the additively encoded SNP genotype, however this time we included the normalized gene expression value of the cis-eQTL gene as a covariate. This should give the relative gene expression change that occurs without the effects of the expression of this gene. We took the regression results and processed with SPIA as described above. For SNPs that act as cis-eQTL for multiple genes, we performed the conditional analysis once for each SNP-gene pair. Similar to the replication analysis, we accepted nominally significant P_{GFDR} values (<0.05) as significant associations.

Functional Annotation

Given the nature of this work, we will focus on the transcription related elements from ENCODE. These functional elements are identified with the following experimental procedures: ChIP-seq of histone proteins and of transcription factors, and DNaseI hypersensitivity assays. ChIP procedures cross-link existing DNA-protein complexes (essentially making the interaction more solid). The long pieces of DNA crossed-linked to the proteins are broken into smaller fragments, and then the fragments of interest are removed for analysis using antibodies to the protein of interest. Following unlinking of the DNA-protein complex of interest, either a

microarray chip (ChIP-chip) or sequencing (ChIP-seq) reveal the sequence of the region being bound by the protein. When targeting histones this gives valuable information on the chromatin state at the region. If transcription factor proteins are targeted, transcription factor binding sites (TFBS) are identified. Although the chromatin state implies information on accessible DNA, DNaseI hypersensitivity assays actually measure regions of open DNA (implying locations of functional elements or active transcription). DNaseI is an enzyme that degrades DNA – in regions of accessible DNA, DNaseI will cleave out multiple fragments. Sequencing of these fragments and alignment to the reference genome allow for identification of these open DNA sites.

Specifically, in this study we will use chromatin state information (enhancers promoters and repressors), regions of open chromatin, and effects of SNPs on regulatory binding motifs found in HaploReg [59]. This database uses information from the HapMap project to identify whether the SNPs of interest are in linkage disequilibrium with these functional elements. Additionally we will examine transcription factor binding sites, and predicted binding locations, along with known gene expression regulators (eQTL) found in RegulomeDB [60]. Many of the cell lines included in ENCODE and regulation information in these databases are from lymphoblastoid cell lines – even in some cases LCLs from the same HapMap samples.

In summary, for each of the SNPs with replicating SNP-Pathway associations, we examined regulatory annotations using HaploReg [59] and RegulomeDB [60]. Specifically we annotated each SNP for:

- Chromatin state information (enhancers promoters and repressors)
- Regions of open chromatin
- Effect of SNPs on regulatory binding motifs
- Transcription factor binding sites (experimentally validated)
- RegulomeDB Functional Score

RegulomeDB scores are based on the level of evidence supporting the regulatory function of the scored variant. For chromatin states, and effects on regulatory binding motifs we did include annotations for variants in strong linkage disequilibrium with our tested SNP. The annotations in question have poor resolution and so using nearby SNPs improves the likelihood of identifying potential regulatory mechanisms.

Annotation of Known SNP and cis-eQTL Gene Effects

We further examined potential explanations underlying replicating SNP-Pathway associations by mapping known associations/functions of both the SNP and its cis-eQTL gene using the NCBI catalog. Specifically we annotated SNP effects using dbSNP [61] and genes using Entrez Gene [62]. We also investigated whether any of the replicating SNPs had known associations from genome-wide association studies. We annotated each SNP using the GWAS catalog [1] and the Johnson GWAS Catalog that contains many suggestive associations [63]. Finally we completed searches of PubMed for each SNP and gene to investigate prior knowledge of SNP and gene function to identify potential explanations for the SNP-pathway association.

CHAPTER IV

RESULTS

A total of 853 genes tested in [14] had one or more cis-eQTL SNPs meeting our threshold. For these 853 genes, a total of 928,908 SNP-gene pairs were test for cis-eQTL activity. After applying a gene-level Bonferroni correction, 22,247 SNP-gene pairs were significant. These results contained 21,315 unique SNPs as many SNPs associated with the expression of multiple genes. The vast majority (20,482) associated with the expression of a single gene. For those associated to multiple genes, 735 associated with two genes, 97 associated with three genes, and 1 associated with four genes. Following quality control procedures, 119 SNPs were removed from analysis as they did not map to the current build of the reference genome or were unavailable in the latest version of HapMap genotyping data. For all remaining SNPs, we filtered out variants in linkage disequilibrium for each gene set, giving a total of 2909 SNPs for future analyses.

Discovery and Replication of Pathway-Based trans-eQTL

In the discovery analysis we tested 2,909 SNPs against 137 KEGG pathways. A total of 291,257 SNP-pathway combinations had at least one differentially expressed gene in the tested pathway. Of these, 240 SNP-Pathway combinations met our Bonferroni-corrected false discovery rates and were carried forward for replication. These results represented a total of 135 SNPs associated with 13 different pathways. Fifty-seven SNPs associated with 2 or more pathways. Of these, 23 SNPs associated with 2 pathways, 24 SNPs associated with 3 pathways, 7 SNPs associated with 4 pathways, 2 SNPs associated with 5 pathways, and 1 SNP associated with 6 pathways. A complete listing of these findings can be found in **Appendix A**.

Of the 135 SNPs significant in the discovery analysis, 65 were available for all HapMap III samples, 65 were available for only a subset of the 1000 Genomes project samples and 5 were unavailable for analysis. Those five SNPs accounted for 19 SNP-pathway combinations. Of the remaining 221 SNP-pathway combinations, 32 met our nominal significance threshold and 15 met our Bonferroni corrected global p-value. Result for both the discovery and replication

Table 1. Significant Replicating SNP-Pathway Associations

SNP	Pathway Name	Discovery						Replication					
		pSize	NDE	pNDE	tA	pPERT	pGFdr	pSize	NDE	pNDE	tA	pPERT	pG
HapMap 3 Samples (n=466)													
rs7586918	Protein proc. in endoplsmc reticulum	105	16	4.56E-09	0.02	0.91	7.66E-06	150	42	3.23E-06	0.57	0.30	1.44E-05
rs10517012	Olfactory transduction	67	10	7.40E-04	-7.36	5.00E-06	8.54E-06	361	17	1.00	6.66	5.00E-06	6.60E-05
rs1162371	Olfactory transduction	67	28	5.93E-06	9.10	5.00E-06	1.00E-07	361	10	1.00	5.52	5.00E-06	6.60E-05
rs11104775	Olfactory transduction	67	36	2.90E-08	11.48	1.00E-03	9.89E-08	361	22	1.00	12.07	5.00E-06	6.60E-05
rs11104947	Olfactory transduction	67	25	1.27E-04	10.36	5.00E-06	1.92E-06	361	17	1.00	10.72	5.00E-06	6.60E-05
1000 Genomes Samples Only (n=236)													
rs6572658	Cell cycle	80	32	8.67E-11	-3.23	0.16	4.78E-08	119	56	7.14E-12	-6.79	0.14	2.84E-11
rs7681425	Parkinson's disease	75	29	1.02E-11	0.19	0.83	2.81E-08	105	53	2.41E-09	-4.08	1.50E-02	9.07E-10
rs7681425	Huntington's disease	120	33	1.15E-08	0.20	0.89	9.16E-06	163	72	7.09E-09	-0.52	0.56	8.12E-08
rs11008749	Cell cycle	80	22	9.51E-09	-2.10	0.34	7.95E-06	119	46	2.25E-08	4.43	0.32	1.41E-07
rs7972875	Huntington's disease	120	44	4.16E-10	-0.28	0.49	6.32E-07	163	42	3.19E-08	-0.24	0.72	4.27E-07
rs12475079	Huntington's disease	120	39	7.02E-09	0.23	0.79	7.23E-06	163	40	3.37E-08	-0.38	0.75	4.64E-07
rs7867279	Epstein-Barr virus inf.	128	54	1.40E-07	-1.51	0.10	8.78E-06	177	52	0.03	5.54	5.00E-06	2.58E-06
rs10131614	RNA trnsprt	84	47	1.60E-11	0.05	0.89	5.07E-08	126	41	4.62E-07	-0.04	0.82	6.01E-06
rs425437	Huntington's disease	120	33	8.34E-10	-0.25	0.73	5.91E-07	163	24	2.53E-06	0.00	1.00	3.51E-05
rs7681425	Alzheimer's disease	101	29	3.10E-08	0.69	0.37	9.16E-06	148	60	4.04E-06	-0.03	0.99	5.37E-05

statistics for the fifteen SNPs passing our Bonferroni correction are presented in **Table 1**. This table contains a few descriptive factors for each association:

- pSize – the number of genes in the pathway with expression measurements
- NDE – the number of differentially expressed genes in the pathway
- pNDE – the probability of observing the number of differentially expressed genes by chance alone
- tA – the total accumulation of expression change through the pathway
- pPERT – the probability of observing that level of accumulation by chance alone
- pGFDR – the false discovery rate corrected combined probability of pNDE and pPERT
- pG – the unadjusted combined probability of pNDE and pPERT

Additionally the table is broken down by origin of the genotype information as SNPs measured by HapMap 3 have nearly double the number of individuals tested as those from the 1000 Genomes Project. A complete listing of nominally significant replication results can be found in **Appendix B**.

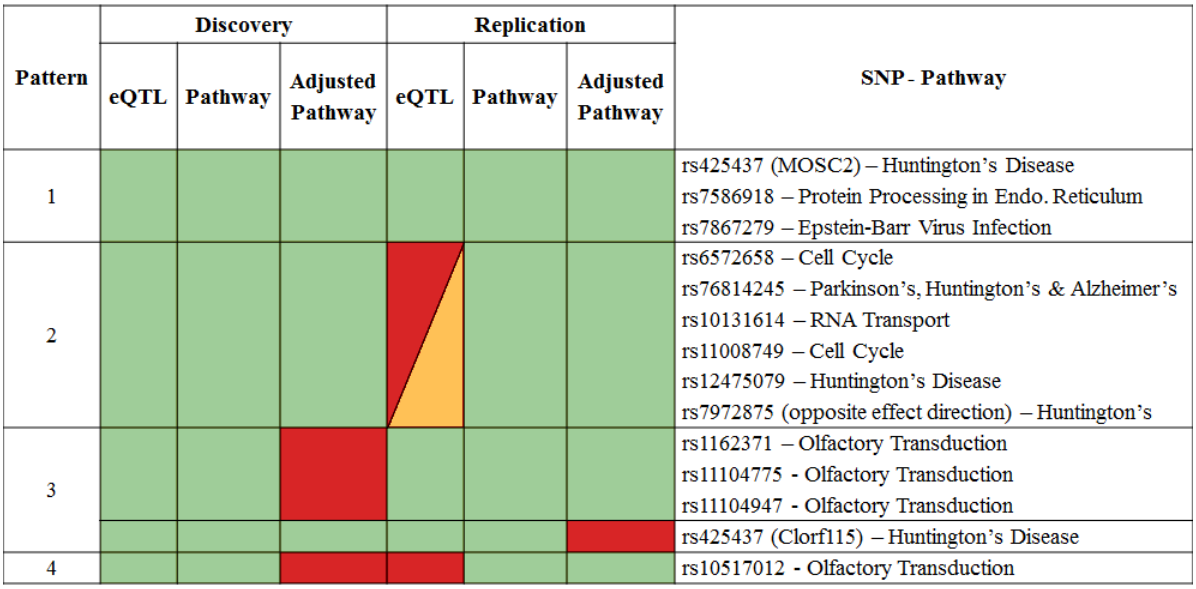


Figure 5. Summary of Effect of Cis-eQTL and Cis-Gene on SNP-Pathway Association. For both the populations we have three types of data: the cis-eQTL, our SNP-pathway association, and the association adjusted for expression level of the cis-gene. Green coloring indicates a statistically significant result, while red represents none significant associations. The yellow component represents the singular case where the cis-eQTL association for rs7972875 is significant but in the opposite direction of effect. From these labels we can group SNPs into different patterns with different interpretations on potential effect mechanism.

Table 2. Cis-gene Adjusted SNP-Pathway Associations

Cis-eQTL	Pathway Name	Population	pSize	NDE	pNDE	tA	pPERT	pG
Cis-eQTL/Gene Pattern 1								
rs425437 MOSC2	Huntington's disease	Discovery	120	56	7.76E-09	-0.03	0.87	1.34E-07
		Replication	163	55	2.44E-04	0.32	0.47	1.15E-03
rs7586918 DTNB	Protein processing in endoplasmic reticulum	Discovery	105	65	7.63E-07	0.60	0.28	3.44E-06
		Replication	150	135	6.88E-13	-0.22	0.78	1.58E-11
rs7867279 ZNF510	Epstein-Barr virus inf.	Discovery	128	92	2.36E-06	-0.17	0.80	2.67E-05
		Replication	177	160	1.83E-13	3.42	0.02	1.31E-13
Cis-eQTL/Gene Pattern 2								
rs6572658 L2HGDH	Cell cycle	Discovery	80	66	3.02E-06	-1.56	0.29	1.32E-05
		Replication	119	107	2.13E-12	-10.17	0.01	9.60E-13
rs7681425 STIM2	Parkinson's disease	Discovery	75	51	9.32E-07	-0.24	0.73	1.03E-05
		Replication	105	79	1.95E-11	-4.30	9.00E-03	5.32E-12
	Huntington's disease	Discovery	120	70	4.01E-05	-0.10	0.61	2.84E-04
		Replication	163	116	3.07E-13	-0.34	0.71	6.58E-12
	Alzheimer's disease	Discovery	101	58	3.18E-04	0.50	0.33	1.06E-03
		Replication	148	92	2.13E-06	-1.02	0.48	1.50E-05
rs10131614 EIF2S1	RNA transport	Discovery	84	73	3.23E-14	-0.01	0.97	1.00E-12
		Replication	126	121	7.42E-18	-0.11	0.77	2.34E-16
rs11008749 ARHGAP12	Cell cycle	Discovery	80	68	1.62E-05	-3.06	0.25	5.34E-05
		Replication	119	104	6.05E-10	10.78	6.00E-03	9.93E-11
rs12475079 IMMT	Huntington's disease	Discovery	120	100	2.66E-06	-0.29	0.22	8.88E-06
		Replication	163	133	4.99E-08	-1.02	0.25	2.37E-07
rs7972875 VPS33A	Huntington's disease	Discovery	120	98	2.98E-07	-0.12	0.60	2.94E-06
		Replication	163	134	2.62E-16	-1.04	6.80E-02	7.04E-16
Cis-eQTL/Gene Pattern 3								
rs1162371 CEP290	Olfactory transduction	Discovery	67	43	0.09	2.38	0.23	0.11
		Replication	361	205	1.00	-15.56	5.00E-06	6.60E-05
rs11104775 CEP290	Olfactory transduction	Discovery	67	47	0.36	-1.04	0.63	0.56
		Replication	361	220	1.00	-59.93	5.00E-06	6.60E-05
rs11104947 CEP290	Olfactory transduction	Discovery	67	44	0.37	2.34	0.33	0.38
		Replication	361	209	1.00	-34.49	5.00E-06	6.60E-05
rs425437 C1orf115	Huntington's disease	Discovery	120	48	1.96E-06	-0.01	0.96	2.67E-05
		Replication	163	20	0.26	0.01	0.48	0.39
Cis-eQTL/Gene Pattern 4								
rs10517012 TMEM33	Olfactory transduction	Discovery	67	24	1.00	1.80	0.27	0.62
		Replication	361	205	1.00	24.88	5.00E-06	6.60E-05

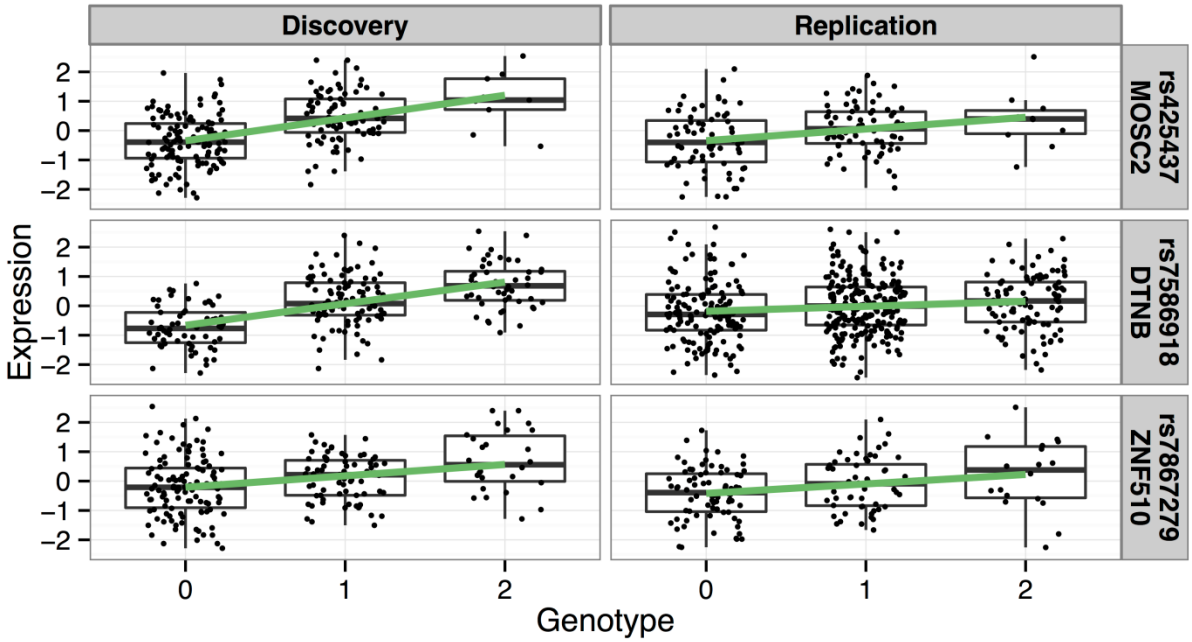


Figure 6. Cis-eQTL Associations for SNPs with Effect Pattern 1. Genotypes are coded based on the number of copies of the minor allele (with respect to the discovery population). Regression lines are presented summarizing direction of effect. Green regression lines indicate associations meeting our p-value threshold while red are not statistically significant.

Identification of Potential Mechanisms of Action

For each SNP-Pathway association we identified the cis-eQTL gene for which the SNP was originally selected. One SNP, rs425437, was a cis-eQTL for two genes, MOSC2 and Clorf115 (open reading frame). Both genes were tested independently in these analyses. Following testing of both cis-eQTL effect and adjustment for effect of cis-gene expression on the SNP-pathway association, a number of combinations/patterns of effects emerged. A summary of the types of patterns observed are presented in **Figure 5**. In this figure, green boxes represent significant associations, while red boxes represent associations that were not statistically significant. Importantly, the SNP with two cis-eQTL, rs425437, had divergent patterns based on each eQTL gene. Also one SNP, rs7972875, had a significant cis-eQTL in the replication cohort, but the opposite direction of effect and was therefore grouped with the other associations that lacked cis-eQTL significance in the replication population.

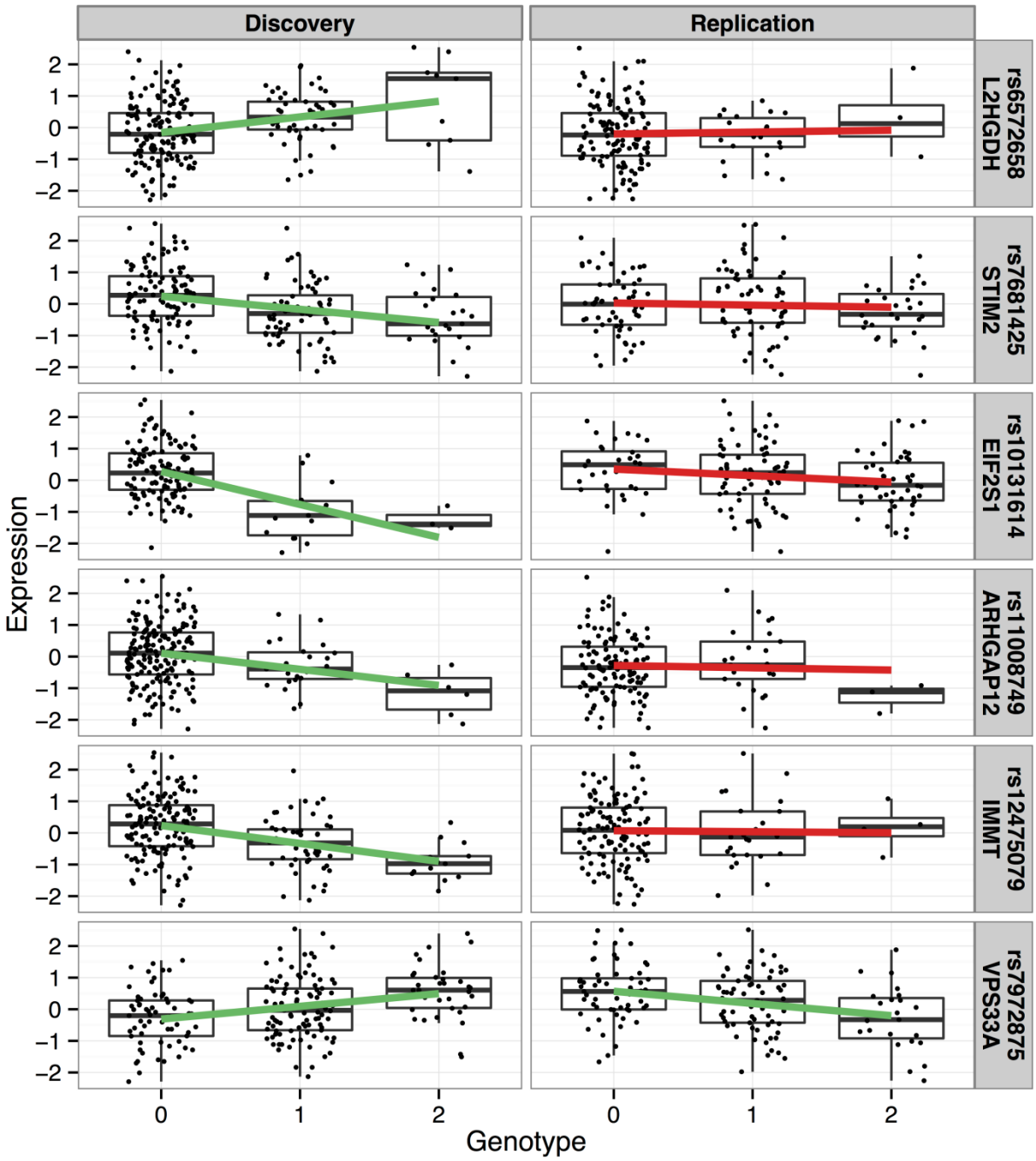


Figure 7. Cis-eQTL Associations for SNPs with Effect Pattern 2. Genotypes are coded based on the number of copies of the minor allele (with respect to the discovery population). Regression lines are presented summarizing direction of effect. Green regression lines indicate associations meeting our p-value threshold while red are not statistically significant.

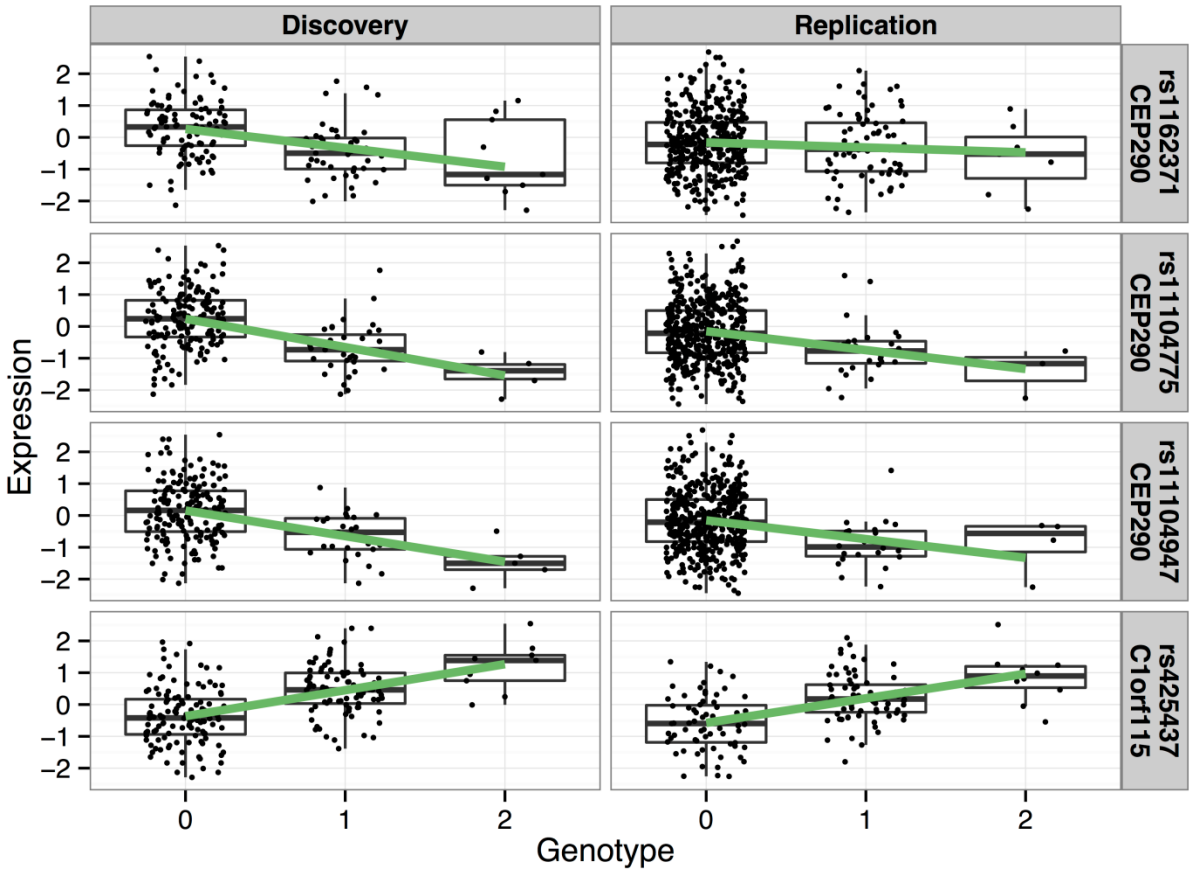


Figure 8. Cis-eQTL Associations for SNPs with Effect Pattern 3. Genotypes are coded based on the number of copies of the minor allele (with respect to the discovery population). Regression lines are presented summarizing direction of effect. Green regression lines indicate associations meeting our p-value threshold while red are not statistically significant.

The results of the SNP-Pathways associations adjusted for the expression level of the cis-gene, are presented in **Table 2**. SNP-gene-pathway combinations are grouped according to the effect pattern described in **Figure 5**. The same statistical measures presented in the unadjusted analyses are included in this table. **Figures 6-9** are plots of each cis-eQTL in the discovery and replication cohorts grouped by their respective effect pattern. Each diagram includes a scatter plot of individual expression measurements by genotype, the mean, first and third quantiles for expression by genotype (numbers are count of minor allele copies), as well a linear regression line. The colors of the line indicate statistical significance based on the gene-specific Bonferroni correction in the discovery population and nominal p-value in the replication population.

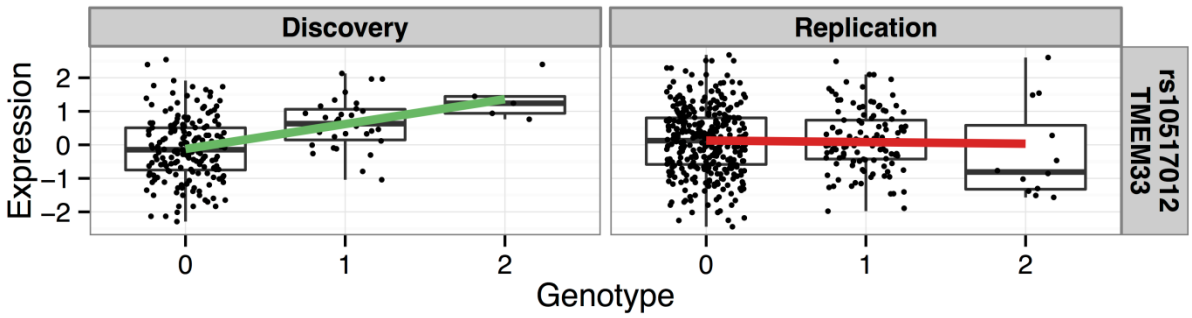


Figure 9. Cis-eQTL Associations for SNPs with Effect Pattern 4. Genotypes are coded based on the number of copies of the minor allele (with respect to the discovery population). Regression lines are presented summarizing direction of effect. Green regression lines indicate associations meeting our p-value threshold while red are not statistically significant.

Complete results of the functional annotation of variants using HaploReg and RegulomeDB are available in **Appendix C**. We curated these results based on relevance to the pathway implicated and the cis-eQTL gene. These filtered results are presented in **Table 3** again grouped by effect pattern. This table contains a variety of information. DNaseI sensitivity reports cell line types for which the SNP is found in open chromatin. Chromatin State includes the interpreted chromatin function for the given cell type/s (in parentheses). Transcription Factor Binding Sites (TFBS) annotate the transcription factor and the cell line tested while Altered Regulatory Motif indicates the relative affinity level of the given transcription factor for the minor allele compared to the reference allele. The final data presented is the RegulomeDB score. This score is determined based on the level of evidence supporting regulatory function. A score of "1f" indicates evidence for an eQTL and either transcription factor binding or a DNase hypersensitivity peak. A score of "5" only requires either transcription factor binding or a DNase hypersensitivity peak at the variant. Scores of 6 are used as an "other" category that indicates minimal binding evidence. Given that these results are only a subset of all the annotations for each SNP, there are some types of data available that are not relevant to the SNP-pathway association. In these cases the boxes are left blank with no background shading. However, when annotation types for variants are not available at all, the entry is shaded with a grey diagonal pattern.

It is important to note that the annotations presented in both **Appendix C** and **Table 3** are only for the listed SNP. However, for annotations from HaploReg we did search on all SNPs in

Table 3. Relevant SNP Functional Annotations

SNP / Gene Pathway	DNaseI Sensitivity	Chromatin State	TFBS	Altered Regulatory Motif	Regulome DB Score
Cis-eQTL/Gene Pattern 1					
rs425437 / MOSC2 Huntington's Disease	• LCL • Glioblastoma	Enhancer (temporal lobe, angular gyrus)		Increased affinity for HIF1	1f
rs7586918 / DTNB Protein processing in ER	• LCL	Strong Enhancer (LCL)			1f
rs7867279 / ZNF510 Epstein-Barr virus inf.	• RPEC			Reduced affinity for NF-1	5
Cis-eQTL/Gene Pattern 2					
rs6572658 / L2HGDH Cell cycle				Increased affinity for EVI-1 and HDAC-2	No Data
rs7681425 / STIM2 Hunt., Park., Alz. Disease	• Prostate adeno-carcinoma	Weak Enhancer (adult liver)			5
rs10131614 / EIF2S1 RNA transport				Reduced affinity for Nanog, and SOX2	6
rs11008749 / ARHGAP12 Cell cycle	• Epidermal keratinocytes			Increased affinity for OCT-1	1f
rs12475079 / IMMT Huntington's disease	• Choroid plexus epithelial cells	Active enhancer (anterior caudate)	• CTCF (brain, muscle)	Reduced affinity for LUN1s	1f
rs7972875 / VPS33A Huntington's disease	• LCL			Reduced affinity for NF-1	5
Cis-eQTL/Gene Pattern 3					
rs1162371 / CEP290 Olfactory transduction		Weak enhancer (cortex)		Reduced affinity for STAT3	6
rs11104775 / CEP290 Olfactory transduction				Reduced affinity for Dlx2	6
rs11104947 / CEP290 Olfactory transduction					6
rs425437 / C1orf115	See entry for rs425437 in Pattern 1.				
Cis-eQTL/Gene Pattern 4					
rs10517012 / TMEM33 Olfactory transduction				Reduced affinity for HNF4	6

linkage disequilibrium with our variant ($R^2 > 0.8$, European descent population). In only one variant did a relevant coding SNP emerge. For rs11104775, cis-eQTL for CEP290 is in moderate LD ($R^2 = 0.81$, European Descent Population) with rs79705698 a missense variant (Asp→Gly) in CEP290. This variant has been deposited into ClinVar – a repository of variants used/discovered in clinical genomic testing – by two individuals. Neither group described the patient's phenotype. Only one provided an assessment of pathogenicity and labeled it as a benign variant.

Following annotation for known phenotypic associations using the GWAS Catalog other GWAS results, only one SNP had a direct pleiotropic association: rs11104947 (Olfactory Transduction). This variant was also associated with vitiligo in a Taiwanese population [64]. Two more variants, rs7972875 and rs12475079 were in LD with variants associated with other disorders. The first SNP, rs7972875, is in moderate LD R^2=0.87) with rs11058789 which is associated to Type II diabetes [65]. The second SNP, rs12475079, is in high LD (R^2= 0.96, and 1.0) with two variants (rs715334 and rs4422155) associated to Parkinson's disease [66] . It is also in high LD ($R^2 = 0.961$) to a variant associated with both Rheumatoid Arthritis and Parkinson's disease [65].

CHAPTER V

DISCUSSION

This study investigated the hypothesis that a single SNP may affect expression levels of distant genes by acting through a biological pathway. To test this hypothesis we performed signaling pathway impact analysis of SNPs known to be cis-eQTL on two independent, multi-ethnic populations. In total we identified 15 highly significant replicating SNP-pathway associations.

Interpretation of Possible Mechanisms of Action

Given our requirement that all SNPs needed to be cis-eQTL, it was possible that the expression of the cis-eQTL gene could be acting as a confounding variable. In this scenario the association between the SNP and pathway may be in fact solely due to effect of the SNP on the cis-gene and the cis-gene effect on the pathway. This confounding possibility is shown in **Figure 10**. We can test the effect of the potential confounder statistically by adjusting our model for expression levels of the cis-gene. Additionally, given that our SNPs were selected as cis-eQTL in our discovery population it was possible that some of those cis-effects may not have replicated in our second population. This natural experimental condition removes the biological impact of the SNP on the gene (and therefore removes the confounding pathway). After testing both scenarios, four combinations of these effects were identified. Each has its own interpretation and as such will be covered individually.

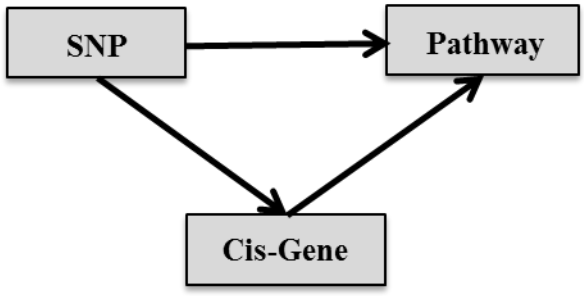


Figure 10. Example of Association Confounding. This figure shows that although a SNP may be associated with expression of a pathway (indicated by an arrow) it may actually be associating through a common factor – the cis-gene.

Cis-eQTL/Gene Pattern 1

In the first pattern, SNP-pathway associations in both the discovery and replication population were robust to expression of the cis-gene when controlled for statistically. Visually this is

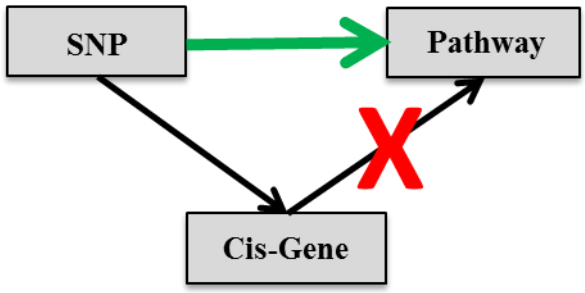


Figure 11. Removal of Confounding by Statistical Correction. This diagram shows that by controlling for expression levels of the cis-gene, we remove the effect of that gene on the pathway and therefore remove the confounding effect. If the pathway is still associated we are measuring an effect independent of the cis-gene (shown as green arrow).

represented in relation to the confounding effect in **Figure 11**. In this group of SNPs, the eQTL did replicate so we are unable to conclude the impact of removal of the cis-eQTL on the SNP-pathway association. However, given our statistical removal of the cis-gene effect we hypothesize that loss of the cis-eQTL would not impact the association.

Cis-eQTL/Gene Pattern 2

In the second pattern, SNP-pathway associations in both the discovery and replication population were robust to expression of the cis-gene when controlled for statistically. Additionally, the cis-eQTL did not replicate in the replication population. Given that the SNP-pathway association replicated in the unadjusted analysis we can conclude that the cis-eQTL effect is not driving for the SNP-pathway association. A summary of this result is shown in **Figure 12.**

Cis-eQTL/Gene Pattern 3

In the third pattern, the SNP-pathway association was not robust to removal of the cis-gene expression in only one population. **Figure 13** displays this effect. If this were observed in both populations there would be stronger evidence to suggest that the SNP-pathway associate is being mediated by the expression of the cis-gene. However, given that we only see this in a single population at a time, a different explanation is needed. Unfortunately

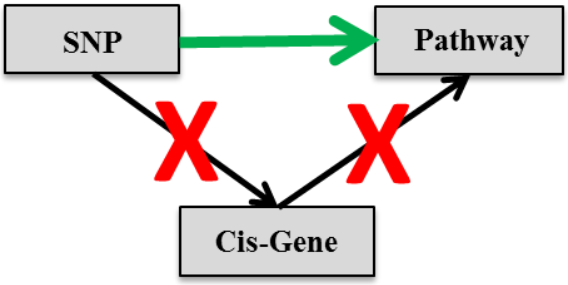


Figure 12. Removal of Confounding by Statistical and Biological Correction. This example shows the scenario where the SNP no longer is associated with the cis-gene (i.e. the eQTL fails to replicate) and where the expression of the cis-gene has been adjusted for statistically. This indicates a SNP-pathway association independent of the cis-gene (shown as green arrow).

without more experimentation it is not clear what these results precisely indicate.

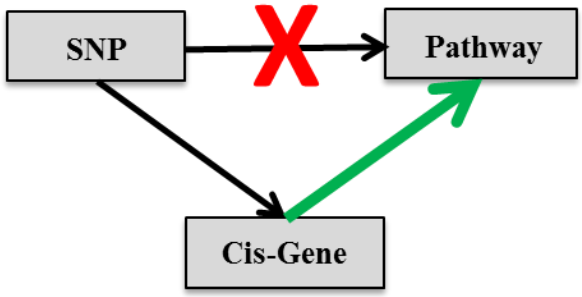


Figure 13. Example of Confounder Driving Association. In this example, when the expression of the cis-gene was accounted for the statistical model, the SNP-pathway association was no longer significant (red X), indicating that the cis-gene is somehow mediating the association (green arrow).

Cis-eQTL/Gene Pattern 4

The fourth pattern is the most difficult to interpret. For this SNP, the eQTL does not replicate in the replication population, additionally adjustment for expression of the cis-gene does not affect the SNP-pathway association (as previously shown in **Figure 12**). However, in the discovery population, adjusting for expression of the cis-gene does remove the SNP-pathway association (**Figure 13**). It is hard to reconcile these two results; it is possible that there is some other, unmeasured factor/s, that are confounding the association between the SNP and the pathway. These factors may be different between our two populations thereby explaining the difference in observed effect.

Plausible Biological Interpretations of Functional Annotations

A number of the pathways identified were associated to multiple SNPs. Ignoring biological heterogeneity, we assume that similar mechanisms underlie each pathway type. For that reason we will discuss the relevant SNP functional annotations and possible mechanisms of action for each pathway individually. Unfortunately for one association, Alzheimer's disease, there were no relevant annotations found for the associated SNP (rs7681425).

Cell Cycle Pathway

Two SNPs, rs6572658 and rs11008749, were associated with the Cell Cycle KEGG pathway. While this is a fairly broad phenotype, there were a number of functional elements for these SNPs that support this association. First, it was found through DNaseI sensitivity assays that the chromatin region around rs11008749 was open in epidermal keratinocytes. This cell type

is the outermost layer of the skin and has very unique cell cycle patterns that are not fully understood [67]. Although both SNPs have annotations for predicted chromatin state, none were particularly compelling for the given pathway. However, both SNPs alter interesting and potentially relevant regulatory motifs. The first SNP, rs6572658, is predicted to have increased affinity for EVI-1 a known oncogene [68]. This SNP also has increased affinity for HDAC-2 which, when abnormally regulated, has been shown to deregulate expression important cell cycle proteins [69]. Finally, rs11008749 has increased affinity for the OCT-1 transcription factor. OCT-1 is required for arresting the cell cycle in the G1 phase of mitosis [70].

Epstein-Barr Virus Infection Pathway

Only one SNP, rs7867279 was associated with the Epstein-Barr Virus (EBV) infection pathway. Interesting the region surrounding this SNP is in open chromatin in retinal pigment epithelial cells (RPEC). This cell type has been shown to not be easily infected by EBV [71]. Additionally, this SNP has reduced affinity for nuclear factor 1 (NF-1). In HeLa cells is was found that a distal NF-1 consensus site enhanced known promoters responsible for triggering the replicative cycle of EBV [72]. It is conceivable that part of the reduced susceptibility of RPEC cells to EBV infection may be related to altered function of NF-1 regulatory regions.

Huntington's Disease Pathway

This pathway was associated with four different SNPs: rs425437, rs7681425, rs12475079, and 7972875. The region surrounding rs425437 and rs7972875 is open chromatin in multiple lymphoblastoid cell lines (LCL) from our HapMap population. Additionally rs7586918 is in a strong enhancer in these cell lines. Rs425437 is in an enhancer in temporal lobe and angular gyrus tissues. Huntington's patients have been found to have significant loss of neurons in the angular gyrus [73,74]. Other chromatin state annotations include rs7681425 in a weak enhancer in liver tissue. Interestingly, mouse models of Huntington's have been observed to have dysfunctional hepatic transcription factors [75].

The most interesting result, however, is rs12475079. This SNP is in open chromatin in choroid plexus epithelial cells. In mouse models of Huntington's disease transplants of choroid plexus epithelial cells have been found to protect against neuron damage. Phenotypically the rats displayed fewer defects in motor function compared to the control animals [76]. Additionally, in

the anterior caudate, rs12475079 is in an active enhancer region. Previous imaging studies have identified concentrated deceased in grey matter in the anterior caudate region for individuals affected by Huntington's. The severity of this atrophy was significantly associated with the number of CAG repeats each patient inherited [77]. It had been hypothesized that the transcription factor CTCF may impact the number of repeats seen in Huntington's as it is associated with many unstable repeat loci. While this was not found in two fibroblast cultures from Huntington's patients, it is possible that in brain tissue this transcription factor may be involved in Huntington's pathogenesis [78]. This background becomes more interesting in light of the fact that rs12475079 is in a CTCF binding site in both brain and muscle tissues.

Olfactory Transduction Pathway

Olfactory transduction was also associated to four SNPs: rs1162371, rs11104775, rs11104947, and rs10517012. The first three of these variants are cis-eQTL for CEP290 – centrosome protein 290kDA. This gene plays a crucial role in the function of cilia and is associated with numerous ciliopathies [79,80]. Malformations of cilia often impact sensory systems, for example, patients with CEP290 mutations causing Leber congenital amaurosis exhibited severely abnormal olfactory function [81]. One SNP, rs11104775, is in moderate LD with a missense variant in CEP290, though it has not been associated to a particular phenotype. While the association may be driven by the missense variant, this SNP also alters a regulatory motif for Dlx2 reducing affinity for this transcription factor. Dlx2 is essentially required for neurogenesis of all olfactory bulb interneurons [82]. One other interesting altered regulatory motif is rs1162371 and STAT3. Phosphorylated STAT3 has been associated with olfactory neuroblastomas, but is not typically observed in normal olfactory tissue [83].

Parkinson's Disease Pathway

Only one SNP was associated with Parkinson's disease, rs7681425. While the most interesting annotations for this variant are for its association to Huntington's disease, this SNP is in open chromatin regions in prostate carcinoma tissue. This is interesting because in a large pedigree study, it was observed that there is a high co-occurrence of prostate cancer and Parkinson's disease. However there has been some indication that this co-occurrence may be due

to drug side effects of certain treatments for Parkinson's symptoms [84]. Our results could perhaps provide an alternative explanation.

Protein Processing in the Endoplasmic Reticulum Pathway

The SNP associated with protein processing in the ER, rs7586918, is in open chromatin in HapMap lymphoblastoid cell lines and is in a strong enhancer region. While there are numerous transcription factors that bind this region, it is not clear how these transcription factors could be specifically related to this very broad pathway.

RNA Transport Pathway

One SNP, rs10131614, was associated with RNA transport. While there is not much functional data available for this variant, it is predicted to alter regulatory motifs and reduce affinity for Nanog and SOX2 transcription factors. Both transcription factors have been shown to interact with and regulate long noncoding RNAs in human cells [85,86].

Limitations

Paradoxically, one of this study's strength is also a significant limitation. The use of multiethnic populations in both the discovery and replication assures for generalizability of results, but also reduced our power to detect true associations. Many associations that are population specific would not be identified in this analysis as we used not only multiethnic populations, but also we used a different mix of populations in the discovery and replication populations. This reduction in power is only amplified by the required expression normalization procedure that allowed us to perform this analysis. As described in the background, this normalization reduces variance in the measured gene expression values. This transformation also limits the interpretation of the gene expression values in our analyses as they are not raw measurements, but rather standardized values without meaning (for instance it is hard to interpret the meaning of a negative gene expression value).

Another limitation is the use of gene expression measurements from lymphoblastoid cell lines. The cell lines have been immortalized, changing their basic cellular properties in the process. While they are widely used is these types of analyses (and results from studies using

these cell lines do generalize to tissue), it is important to recognize that these are not measurements of a natural environment. The final limitation is the use of the KEGG database. While this too is commonly used, due to changes in licensing, this resource has not been updated in a number of years. While there is no evidence that the source is inaccurate, if the resource were to be updated it likely would contain more (and more detailed) information.

Conclusions and Future Directions

This study identified 32 SNP-Pathway associations that replicated across multiple ethnic cohorts. Fifteen of these SNP-Pathway associations were especially robust and were investigated more deeply. In these 15 associations, 4 potential patterns of action with respect to cis-eQTL function were identified and interpreted. Finally, functional annotation provided further insight into the validity and possible mechanism of action underlying these associations. Future work should try to replicate these results in primary tissue samples and investigate potential phenotypic associations of the variants identified by this approach.

REFERENCES

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS catalog, a curated resource of snp-trait associations. Nucleic Acids Res 2014, Jan;**42**(Database issue):D1001-6.

2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences 2009;**106**(23):9362-7.

3. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from GWAS. PLoS Genet 2010, Apr;**6**(4):e1000888.

4. Gibson G, Weir B. The quantitative genetics of transcription. Trends Genet 2005, Nov;**21**(11):616-23.

5. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, et al. Heritability and tissue specificity of expression quantitative trait loci. PLoS Genet 2006, Oct 20;**2**(10):e172.

6. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet 2007, Oct;**39**(10):1217-24.

7. Duggal G, Wang H, Kingsford C. Higher-order chromatin domains link eqtls with the expression of far-away genes. Nucleic Acids Res 2014, Jan;**42**(1):87-96.

8. Phillips J, Eberwine JH. Antisense RNA amplification: A linear amplification method for analyzing the mrna population from single living cells. Methods 1996;**10**(3):283-8.

9. Lu Q, Richardson B. DNaseI hypersensitivity analysis of chromatin structure. Methods Mol Biol 2004;**287**:77-86.

10. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 2010, Aug;**28**(8):817-25.

11. Fuda NJ, Ardehali MB, Lis JT. Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature 2009, Sep 10;**461**(7261):186-92.

12. Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, et al. Heritability of alternative splicing in the human genome. Genome Res 2007, Aug;**17**(8):1210-8.

13. Yang S, Liu Y, Jiang N, Chen J, Leach L, Luo Z, Wang M. Genome-wide eqtls and heritability for gene expression traits in unrelated individuals. BMC Genomics 2014;**15**:13.

14. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-qtls yields insight into human gene regulation. PLoS Genet 2008, Oct;**4**(10):e1000214.

15. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet 2012, Apr;**8**(4):e1002639.

16. Michaelson JJ, Loguercio S, Beyer A. Detection and interpretation of expression quantitative trait loci (eqtl). Methods 2009, Jul;**48**(3):265-76.

17. McKenzie M, Henders AK, Caracella A, Wray NR, Powell JE. Overlap of expression quantitative trait loci (eqtl) in human brain and blood. BMC Med Genomics 2014, Jun 3;**7**(1):31.

18. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet 2011, May;**7**(5):e1002078.

19. Hsiao CL, Lian IeB, Hsieh AR, Fann CS. Modeling expression quantitative trait loci in data combining ethnic populations. BMC Bioinformatics 2010;**11**:111.

20. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eqtls as putative drivers of known disease associations. Nat Genet 2013, Sep;**45**(10):1238.

21. Sudarsanam P, Cohen BA. Single nucleotide variants in transcription factors associate more tightly with phenotype than with gene expression. PLoS Genet 2014, May;**10**(5):e1004325.

22. Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet 2011, Nov;**43**(11):1082-90.

23. Mangravite LM, Engelhardt BE, Medina MW, Smith JD, Brown CD, Chasman DI, et al. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. Nature 2013, Oct 17;**502**(7471):377-80.

24. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, et al. Genetical genomics: Spotlight on QTL hotspots. PLoS Genet 2008, Oct;**4**(10):e1000232.

25. Gaffney DJ. Global properties and functional complexity of human gene regulatory variation. PLoS Genet 2013, May;**9**(5):e1003501.

26. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 2012, Oct;**44**(10):1084-9.

27. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet 2011, Aug;**7**(8):e1002197.

28. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eqtls as putative drivers of known disease associations. Nat Genet 2013, Oct;**45**(10):1238-43.

29. Williams RW. Expression genetics and the phenotype revolution. Mammalian Genome 2006;**17**(6):496-502.

30. West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St Clair DA. Global eqtl mapping reveals the complex genetic architecture of transcript-level variation in arabidopsis. Genetics 2007, Mar;**175**(3):1441-50.

31. Brynedal B, Raj T, Stranger BE, Bjornson R, Neale BM, Voight BF, Cotsapas C. Cross-phenotype meta-analysis reveals large-scale trans-eqtls mediating patterns of transcriptional co-regulation. ArXiv Preprint ArXiv:1402.1728 2014;.

32. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: The promise of eqtl studies. Trends Genet 2008, Aug;**24**(8):408-15.

33. Lee E, Bussemaker HJ. Identifying the genetic determinants of transcription factor activity. Mol Syst Biol 2010, Sep 21;**6**:412.

34. Wu C, Delano DL, Mitro N, Su SV, Janes J, McClurg P, et al. Gene set enrichment in eqtl data identifies novel annotations and pathway regulators. PLoS Genet 2008;**4**(5):e1000070.

35. Wessel J, Zapala MA, Schork NJ. Accommodating pathway information in expression quantitative trait locus analysis. Genomics 2007, Jul;**90**(1):132-42.

36. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T. EQED: An efficient method for interpreting eqtl associations using protein networks. Mol Syst Biol 2008, Mar 4;**4**.

37. Rashid I, McDermott J, Samudrala R. Inferring molecular interactions pathways from eqtl data. Computational Systems Biology 2009, Jan;:211-23.

38. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: Population substructure and genome-wide association studies. Hum Mol Genet 2008, Oct 15;**17**(R2):R143-50.

39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet 2000, May;**25**(1):25-9.

40. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2014, Jan;**42**(Database issue):D472-7.

41. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, et al. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers (Basel) 2012;**4**(4):1180-211.

42. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;**28**(1):27-30.

43. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: Back to metabolism in KEGG. Nucleic Acids Res 2014, Jan;**42**(Database issue):D199-205.

44. Khatri P, Drăghici S. Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics 2005;**21**(18):3587-95.

45. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005, Oct 25;**102**(43):15545-50.

46. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. Bioinformatics 2009, Jan 1;**25**(1):75-82.

47. Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001, Mar;**27**(3):234-6.

48. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The international hapmap project. Nature 2003, Dec;**426**(6968):789.

49. International HapMap Consortium. A haplotype map of the human genome. Nature 2005, Oct 27;**437**(7063):1299-320.

50. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million snps. Nature 2007, Oct 18;**449**(7164):851-61.

51. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010, Sep 2;**467**(7311):52-8.

52. Consortium T1GP. A map of human genome variation from population-scale sequencing. Nature 2010, Oct;**467**(7319):1061.

53. ENCODE Project Consortium. The ENCODE (encyclopedia of DNA elements) project. Science 2004, Oct 22;**306**(5696):636-40.

54. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007, Jun 14;**447**(7146):799-816.

55. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature 2012, Sep 6;**489**(7414):57-74.

56. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010, Dec;**34**(8):816-34.

57. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 2007;**81**(3):559-75.

58. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 2001;:1165-88.

59. Ward LD, Kellis M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012, Jan;**40**(Database issue):D930-4.

60. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using regulomedb. Genome Res 2012, Sep;**22**(9):1790-7.

61. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. DbSNP: The NCBI database of genetic variation. Nucleic Acids Res 2001;**29**(1):308-11.

62. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: Gene-centered information at NCBI. Nucleic Acids Res 2011, Jan;**39**(Database issue):D52-7.

63. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. BMC Med Genet 2009;**10**:6.

64. Lan CC, Ko YC, Tu HP, Wu CS, Lee CH, Wu CS, Yu HS. Association study between keratinocyte-derived growth factor gene polymorphisms and susceptibility to vitiligo vulgaris in a taiwanese population: Potential involvement of stem cell factor. Br J Dermatol 2009, Jun;**160**(6):1180-7.

65. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007, Jun 7;**447**(7145):661-78.

66. Fung HC, Scholz S, Matarin M, Simón-Sánchez J, Hernandez D, Britton A, et al. Genome-wide genotyping in parkinson's disease and neurologically normal controls: First stage analysis and public release of data. Lancet Neurol 2006, Nov;**5**(11):911-6.

67. Gandarillas A, Freije A. Cycling up the epidermis: Reconciling 100 years of debate. Exp Dermatol 2014, Feb;**23**(2):87-91.

68. Buonamici S, Chakraborty S, Senyuk V, Nucifora G. The role of EVI1 in normal and leukemic cells. Blood Cells, Molecules, and Diseases 2003;**31**(2):206-12.

69. Noh JH, Jung KH, Kim JK, Eun JW, Bae HJ, Xie HJ, et al. Aberrant regulation of HDAC2 mediates proliferation of hepatocellular carcinoma cells by deregulating expression of G1/S cell cycle proteins. PLoS One 2011;**6**(11):e28103.

70. Dalvai M, Schubart K, Besson A, Matthias P. Oct1 is required for mtor-induced G1 cell cycle arrest via the control of p27(kip1) expression. Cell Cycle 2010, Oct 1;**9**(19):3933-44.

71. Huemer HP, Larcher C, Kirchebner W, Klingenschmid J, Göttinger W, Irschick EU. Susceptibility of human retinal pigment epithelial cells to different viruses. Graefes Arch Clin Exp Ophthalmol 1996, Mar;**234**(3):177-85.

72. Glaser G, Vogel M, Wolf H, Niller HH. Regulation of the epstein-barr viral immediate early BRLF1 promoter through a distal NF1 site. Arch Virol 1998;**143**(10):1967-83.

73. Macdonald V, Halliday GM, Trent RJ, McCusker EA. Significant loss of pyramidal neurons in the angular gyrus of patients with huntington's disease. Neuropathol Appl Neurobiol 1997, Dec;**23**(6):492-5.

74. Ho AK, Nestor PJ, Williams GB, Bradshaw JL, Sahakian BJ, Robbins TW, Barker RA. Pseudo-neglect in huntington's disease correlates with decreased angular gyrus density. Neuroreport 2004, Apr 29;**15**(6):1061-4.

75. Chiang MC, Chern Y, Juo CG. The dysfunction of hepatic transcriptional factors in mice with huntington's disease. Biochim Biophys Acta 2011, Sep;**1812**(9):1111-20.

76. Borlongan CV, Thanos CG, Skinner SJ, Geaney M, Emerich DF. Transplants of encapsulated rat choroid plexus cells exert neuroprotection in a rodent model of huntington's disease. Cell Transplantation 2007;**16**(10):987-92.

77. Kassubek J, Juengling FD, Kioschies T, Henkel K, Karitzky J, Kramer B, et al. Topography of cerebral atrophy in early huntingtons disease: A voxel based morphometric MRI study. Journal of Neurology, Neurosurgery & Psychiatry 2004;**75**(2):213-20.

78. Pepers BA, Mastrokolias A, van Ommen G-J, den Dunnen JT, Hoen PB, van Roon-Mom WMC. B15 CTCF in huntington's disease. Journal of Neurology, Neurosurgery & Psychiatry 2012;**83**(Suppl 1):A10-.

79. Travaglini L, Brancati F, Attie-Bitach T, Audollent S, Bertini E, Kaplan J, et al. Expanding CEP290 mutational spectrum in ciliopathies. Am J Med Genet A 2009, Oct;**149A**(10):2173-80.

80. Coppieters F, Lefever S, Leroy BP, De Baere E. CEP290, a gene with many faces: Mutation overview and presentation of cep290base. Hum Mutat 2010, Oct;**31**(10):1097-108.

81. McEwen DP, Koenekoop RK, Khanna H, Jenkins PM, Lopez I, Swaroop A, Martens JR. Hypomorphic CEP290/NPHP6 mutations result in anosmia caused by the selective loss of G proteins in cilia of olfactory sensory neurons. Proc Natl Acad Sci U S A 2007, Oct 2;**104**(40):15917-22.

82. Brill MS, Snapyan M, Wohlfrom H, Ninkovic J, Jawerka M, Mastick GS, et al. A dlx2- and pax6-dependent transcriptional code for periglomerular neuron specification in the adult olfactory bulb. J Neurosci 2008, Jun 18;**28**(25):6439-52.

83. Zeng M, Cui Y, Wu C. [Expression of SSTR2 and P-STAT3 in human olfactory neuroblastoma]. Lin Chung Er Bi Yan Hou Tou Jing Wai Ke Za Zhi 2010, Aug;**24**(15):690-2.

84. Kareus SA, Figueroa KP, Cannon-Albright LA, Pulst SM. Shared predispositions of parkinsonism and cancer: A population-based pedigree-linked study. Arch Neurol 2012, Dec;**69**(12):1572-7.

85. Sheik Mohamed J, Gaughwin PM, Lim B, Robson P, Lipovich L. Conserved long noncoding rnas transcriptionally regulated by oct4 and nanog modulate pluripotency in mouse embryonic stem cells. RNA 2010, Feb;**16**(2):324-37.

86. Ng SY, Bogu GK, Soh BS, Stanton LW. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. Mol Cell 2013, Aug 8;**51**(3):349-59.

APPENDIX A

COMPLETE DISCOVERY ASSOCIATION RESULTS

This table contains all of the significant discovery SNP-Pathway association results. For each SNP-pathway combination we report the following measures:

- pSize – The number of genes in the pathway with gene expression values available
- NDE – The number of differentially expressed genes in the pathway
- pNDE – The probability of observing the number of differentially expressed genes by chance alone.
- tA – The total accumulation of effect from differentially expressed genes in the pathway
- pPERT – The probability of observing the total accumulation value by chance alone
- pGFDR – The false discovery rate adjusted combined probability of pNDE and pPERT.

Results are provided in no particular order, though SNPs with associations to multiple pathways have their SNP-Pathway statistics grouped together.

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs10012092	Huntington's disease	120	59	4.00E-12	-0.93	9.70E-02	1.49E-09
rs10012092	Parkinson's disease	75	43	4.63E-12	1.63	2.97E-01	2.53E-09
rs10012092	RNA transport	84	44	1.55E-10	0.87	2.40E-02	4.41E-09
rs1004579	RNA transport	84	31	4.60E-09	0.29	3.68E-01	4.59E-06
rs10131614	RNA transport	84	47	1.60E-11	0.05	8.89E-01	5.07E-08
rs10264186	Alzheimer's disease	101	45	4.13E-16	-1.27	2.66E-01	1.77E-13
rs10264186	Huntington's disease	120	59	1.89E-23	-0.62	3.15E-01	2.08E-20
rs10264186	Parkinson's disease	75	46	6.20E-24	1.99	1.56E-01	6.96E-21
rs10517012	Olfactory transduction	67	10	7.40E-04	-7.36	5.00E-06	8.54E-06
rs1060435	Cell cycle	80	31	2.33E-09	2.31	1.60E-01	5.25E-07
rs1060435	Huntington's disease	120	46	5.00E-13	0.42	4.34E-01	8.11E-10

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs1060435	Parkinson's disease	75	28	3.60E-08	1.44	1.51E-01	4.51E-06
rs1061338	Alzheimer's disease	101	39	5.35E-11	-0.93	3.20E-01	2.04E-08
rs1061338	Huntington's disease	120	44	2.42E-11	-0.23	7.20E-01	2.04E-08
rs1061338	Parkinson's disease	75	37	1.83E-14	0.74	5.74E-01	4.73E-11
rs10788823	Cell cycle	80	23	2.01E-11	-1.97	3.95E-01	2.72E-08
rs10794021	Cell cycle	80	21	8.03E-09	2.46	1.22E-01	2.54E-06
rs10850838	RNA transport	84	39	3.82E-12	-0.57	4.30E-02	6.44E-10
rs10998219	Protein processing in endoplasmic reticulum	105	26	1.67E-12	-2.41	6.50E-02	4.06E-10
rs11008749	Cell cycle	80	22	9.51E-09	-2.10	3.39E-01	7.95E-06
rs11104775	Olfactory transduction	67	36	2.90E-08	11.48	1.00E-03	9.89E-08
rs11104947	Olfactory transduction	67	25	1.27E-04	10.36	5.00E-06	1.92E-06
rs11164929	Cell cycle	80	42	1.92E-11	4.84	2.20E-02	1.64E-09
rs11164929	RNA transport	84	39	9.96E-09	0.11	8.82E-01	1.13E-05
rs11230687	Alzheimer's disease	101	28	1.01E-11	-1.25	3.07E-01	3.38E-09
rs11230687	Huntington's disease	120	35	4.53E-15	-0.40	7.00E-01	6.49E-12
rs11230687	Parkinson's disease	75	29	2.25E-16	-1.26	4.33E-01	4.39E-13
rs1162371	Olfactory transduction	67	28	5.93E-06	9.10	5.00E-06	1.00E-07
rs11704195	Alzheimer's disease	101	32	9.60E-13	2.01	3.71E-01	4.58E-10
rs11704195	Huntington's disease	120	39	1.15E-15	0.74	5.61E-01	1.51E-12
rs11704195	Parkinson's disease	75	31	4.46E-16	-1.93	5.19E-01	1.11E-12
rs1179434	RNA transport	84	36	1.57E-09	-0.02	8.54E-01	3.91E-06
rs11888	Parkinson's disease	75	25	7.16E-12	0.74	4.08E-01	9.74E-09
rs1208077	Huntington's disease	120	51	5.47E-09	-0.38	4.88E-01	3.79E-06
rs1208077	Parkinson's disease	75	40	5.89E-11	-0.32	8.10E-01	1.62E-07

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs12150997	Huntington's disease	120	38	1.67E-08	-0.24	5.54E-01	7.57E-06
rs12150997	Parkinson's disease	75	33	7.25E-12	-0.74	4.10E-01	1.03E-08
rs12150997	RNA transport	84	31	5.97E-09	0.25	3.28E-01	2.60E-06
rs12200420	Protein processing in endoplasmic reticulum	105	41	5.71E-09	-0.85	5.03E-01	8.07E-06
rs12238713	Parkinson's disease	75	47	4.00E-09	2.33	3.16E-01	3.72E-06
rs12274436	RNA transport	84	35	1.06E-08	0.38	5.85E-01	1.66E-05
rs12475079	Huntington's disease	120	39	7.02E-09	0.23	7.85E-01	7.23E-06
rs12475079	RNA transport	84	33	4.13E-10	-0.29	2.91E-01	3.76E-07
rs12511773	Cell cycle	80	45	9.23E-11	-1.97	3.53E-01	3.63E-08
rs12511773	Epstein-Barr virus infection	128	56	1.04E-07	-2.58	1.80E-02	1.05E-06
rs12511773	Huntington's disease	120	60	5.55E-11	-0.61	1.93E-01	1.87E-08
rs12511773	Protein processing in endoplasmic reticulum	105	52	1.63E-09	-0.89	2.80E-01	3.42E-07
rs12511773	RNA transport	84	52	1.49E-14	0.62	3.50E-02	2.51E-12
rs12517057	Cell cycle	80	38	8.15E-14	-5.45	2.70E-02	9.79E-12
rs12517057	RNA transport	84	35	8.13E-11	0.23	7.79E-01	9.92E-08
rs12574149	Protein processing in endoplasmic reticulum	105	26	2.54E-11	-0.53	4.21E-01	3.20E-08
rs1265163	Protein processing in endoplasmic reticulum	105	44	1.37E-14	-0.07	9.29E-01	5.66E-11
rs12800372	Alzheimer's disease	101	43	8.19E-14	-1.46	2.93E-01	5.01E-11
rs12800372	Huntington's disease	120	41	1.17E-09	-1.03	3.90E-02	4.85E-08
rs12800372	Parkinson's disease	75	37	1.51E-14	0.10	9.50E-01	5.01E-11
rs12817892	RNA transport	84	33	4.45E-08	-1.77	5.00E-06	8.92E-10

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs13013390	RNA transport	84	39	1.69E-09	-0.01	9.44E-01	4.54E-06
rs13191247	Protein processing in endoplasmic reticulum	105	38	8.26E-14	0.02	9.79E-01	3.25E-10
rs1378162	Cell cycle	80	31	3.75E-10	2.35	2.86E-01	3.31E-07
rs1388970	Cell cycle	80	50	1.22E-12	2.91	2.07E-01	4.98E-10
rs1388970	RNA transport	84	55	4.83E-15	0.70	7.20E-02	1.67E-12
rs1395259	Mineral absorption	31	6	9.30E-04	-0.34	5.00E-06	1.13E-05
rs1609798	Alzheimer's disease	101	31	3.16E-09	-1.11	1.57E-01	4.71E-07
rs1609798	Huntington's disease	120	41	1.70E-13	-0.27	5.61E-01	1.87E-10
rs1609798	Parkinson's disease	75	33	8.77E-15	0.01	9.90E-01	3.68E-11
rs1634761	RNA transport	84	36	4.20E-11	-0.03	7.24E-01	1.04E-07
rs175006	Cell cycle	80	29	4.25E-09	2.49	2.19E-01	2.70E-06
rs17605444	RNA transport	84	43	5.95E-16	0.02	9.13E-01	2.49E-12
rs17643917	Cell cycle	80	28	3.46E-10	-2.39	2.44E-01	2.68E-07
rs1790807	RNA transport	84	44	5.27E-08	0.64	4.00E-02	6.01E-06
rs1792285	RNA transport	84	32	5.50E-07	-0.98	1.00E-02	1.48E-05
rs1806294	Protein processing in endoplasmic reticulum	105	33	1.59E-19	-0.25	7.62E-01	5.52E-16
rs1878014	Protein processing in endoplasmic reticulum	105	24	1.64E-08	0.59	2.64E-01	1.01E-05
rs1885499	Protein processing in endoplasmic reticulum	105	41	3.64E-08	1.29	9.50E-02	9.57E-06
rs1903262	Huntington's disease	120	32	1.42E-10	-0.01	9.80E-01	3.79E-07
rs1903262	Parkinson's disease	75	24	5.02E-10	0.58	5.51E-01	3.79E-07
rs1947457	Cell cycle	80	36	3.11E-11	3.92	2.35E-01	2.63E-08
rs2073734	Alzheimer's disease	101	35	5.60E-08	3.10	1.76E-01	8.69E-06

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs2073734	Huntington's disease	120	43	5.21E-10	0.55	8.43E-01	6.73E-07
rs2073734	Parkinson's disease	75	35	3.54E-12	-1.70	5.90E-01	7.93E-09
rs2074774	Dopaminergic synapse	78	12	1.38E-03	-1.46	5.00E-06	1.69E-05
rs210280	Cell cycle	80	16	1.84E-09	-3.00	6.60E-02	2.43E-07
rs2114647	Protein processing in endoplasmic reticulum	105	37	1.17E-18	-0.20	8.49E-01	5.54E-15
rs2184334	Protein processing in endoplasmic reticulum	105	25	2.61E-09	-0.13	7.81E-01	5.04E-06
rs2203712	Parkinson's disease	75	21	1.31E-09	0.09	8.01E-01	2.59E-06
rs2239705	Alzheimer's disease	101	44	4.63E-08	-1.15	3.49E-01	1.37E-05
rs2239705	Huntington's disease	120	54	3.26E-10	-0.42	4.95E-01	5.02E-07
rs2239705	Parkinson's disease	75	39	5.19E-10	0.85	6.32E-01	5.02E-07
rs2290507	Alzheimer's disease	101	39	5.23E-11	1.23	2.32E-01	4.22E-08
rs2290507	Parkinson's disease	75	32	1.34E-10	-0.02	9.85E-01	2.09E-07
rs2298581	Protein processing in endoplasmic reticulum	105	19	1.12E-10	0.44	3.73E-01	1.03E-07
rs2303115	Cell cycle	80	24	1.62E-09	-2.84	4.50E-02	2.20E-07
rs2332496	Huntington's disease	120	45	2.95E-08	0.44	1.83E-01	4.83E-06
rs2332496	Parkinson's disease	75	34	5.50E-09	-0.02	9.68E-01	4.83E-06
rs2332496	RNA transport	84	39	1.72E-10	-0.70	2.00E-02	1.26E-08
rs2428521	Olfactory transduction	67	27	6.46E-04	-9.60	5.00E-06	9.09E-06
rs243324	Cell cycle	80	38	9.00E-11	-2.15	1.79E-01	2.69E-08
rs243324	Fanconi anemia pathway	25	17	1.29E-08	0.23	2.55E-01	1.74E-06

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs243324	Huntington's disease	120	54	1.64E-13	-0.23	4.70E-01	3.10E-10
rs243324	Parkinson's disease	75	37	3.91E-11	0.30	6.83E-01	2.91E-08
rs243324	RNA transport	84	38	5.37E-10	0.35	9.30E-02	3.98E-08
rs252646	Cell cycle	80	21	8.10E-12	-3.87	3.30E-02	9.77E-10
rs2526478	Cell cycle	80	30	1.75E-09	-2.21	1.43E-01	3.62E-07
rs2526478	Huntington's disease	120	38	3.44E-09	-0.24	4.51E-01	1.37E-06
rs2526478	RNA transport	84	33	6.29E-11	0.10	9.53E-01	1.84E-07
rs266805	Huntington's disease	120	57	9.95E-12	-0.21	5.50E-01	2.00E-08
rs266805	RNA transport	84	42	7.22E-10	0.51	7.10E-02	8.60E-08
rs2668427	Epstein-Barr virus infection	128	75	1.08E-08	0.72	6.07E-01	5.95E-06
rs2668427	Huntington's disease	120	76	5.02E-11	-0.48	4.02E-01	3.54E-08
rs2668427	Parkinson's disease	75	49	3.13E-08	1.50	2.97E-01	6.21E-06
rs2668427	RNA transport	84	67	9.30E-18	0.74	8.00E-03	4.59E-16
rs2688590	RNA transport	84	49	1.03E-15	-0.11	7.94E-01	3.86E-12
rs2695317	Protein processing in endoplasmic reticulum	105	34	2.41E-18	0.19	8.09E-01	1.08E-14
rs2746029	Huntington's disease	120	30	3.51E-12	0.20	8.04E-01	9.18E-09
rs2746029	Parkinson's disease	75	21	6.65E-10	-0.67	4.01E-01	3.63E-07
rs277384	Protein processing in endoplasmic reticulum	105	39	8.18E-11	-0.27	7.11E-01	1.77E-07
rs277384	RNA transport	84	30	3.52E-08	0.27	3.55E-01	1.49E-05
rs2915228	Viral myocarditis	35	5	1.08E-03	-0.92	5.00E-06	1.17E-05
rs2967359	Alzheimer's disease	101	39	2.56E-09	-0.92	2.04E-01	5.22E-07
rs2967359	Huntington's disease	120	45	4.74E-10	-0.33	4.10E-01	4.73E-07
rs2967359	Parkinson's disease	75	33	8.26E-10	-0.80	3.73E-01	4.73E-07

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs3095250	Alzheimer's disease	101	43	5.82E-14	-0.68	4.66E-01	3.82E-11
rs3095250	Huntington's disease	120	50	1.48E-15	-0.46	2.19E-01	7.78E-13
rs3095250	Parkinson's disease	75	42	3.58E-19	1.22	3.40E-01	7.11E-16
rs329312	Protein processing in endoplasmic reticulum	105	19	3.21E-09	-0.13	7.92E-01	4.60E-06
rs3747956	Cell cycle	80	23	3.90E-09	-2.20	1.00E-01	1.06E-06
rs3750131	RNA transport	84	47	6.02E-10	0.04	8.73E-01	1.57E-06
rs3750132	RNA transport	84	42	3.72E-09	0.14	6.05E-01	6.31E-06
rs3823943	Huntington's disease	120	38	5.05E-10	0.21	5.73E-01	4.45E-07
rs3823943	Parkinson's disease	75	29	2.86E-10	-0.05	9.36E-01	4.45E-07
rs3910384	Parkinson's disease	75	26	2.18E-09	-0.40	7.42E-01	4.39E-06
rs4144887	Cell cycle	80	38	1.84E-12	-5.87	7.00E-03	2.86E-11
rs4144887	RNA transport	84	40	4.31E-13	0.75	1.60E-02	2.86E-11
rs425437	Alzheimer's disease	101	29	2.97E-09	-1.06	2.00E-01	5.91E-07
rs425437	Huntington's disease	120	33	8.34E-10	-0.25	7.25E-01	5.91E-07
rs425437	Parkinson's disease	75	27	3.37E-11	-1.38	2.23E-01	2.64E-08
rs4281907	RNA transport	84	42	4.73E-09	-0.29	2.42E-01	3.28E-06
rs4346637	Alzheimer's disease	101	44	1.35E-13	-1.00	2.22E-01	4.27E-11
rs4346637	Huntington's disease	120	56	1.22E-18	-0.52	2.28E-01	1.62E-15
rs4346637	Parkinson's disease	75	41	4.04E-17	-1.78	9.70E-02	1.07E-14
rs4346637	RNA transport	84	30	2.43E-07	0.76	7.00E-03	1.20E-06
rs4489748	Cell cycle	80	28	1.59E-11	0.33	8.88E-01	4.54E-08
rs4626725	Cell cycle	80	58	3.06E-10	-2.81	3.68E-01	1.84E-07
rs4626725	RNA transport	84	61	9.46E-11	0.89	1.90E-02	6.91E-09
rs4674297	Alzheimer's disease	101	48	7.33E-09	2.30	5.10E-02	2.82E-07
rs4674297	Huntington's disease	120	61	1.79E-12	0.43	3.24E-01	1.12E-09

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs4674297	Parkinson's disease	75	44	3.91E-12	0.36	7.75E-01	3.70E-09
rs4674297	RNA transport	84	53	3.00E-16	-0.86	1.20E-02	1.97E-14
rs4750935	Alzheimer's disease	101	46	1.79E-11	-2.30	1.03E-01	2.29E-09
rs4750935	Huntington's disease	120	60	6.19E-17	-0.40	5.97E-01	9.55E-14
rs4750935	Parkinson's disease	75	46	6.80E-18	0.90	6.24E-01	2.31E-14
rs4750935	RNA transport	84	34	2.70E-07	1.37	4.00E-03	7.77E-07
rs4899667	RNA transport	84	27	3.60E-09	-0.02	9.61E-01	9.14E-06
rs5743030	Cell cycle	80	31	5.75E-13	1.53	5.71E-01	1.22E-09
rs6073555	Alzheimer's disease	101	41	1.95E-08	-1.01	4.03E-01	6.83E-06
rs6073555	Huntington's disease	120	47	7.16E-09	0.06	8.96E-01	6.83E-06
rs6073555	Parkinson's disease	75	39	4.87E-12	-0.55	7.68E-01	1.36E-08
rs6075348	RNA transport	84	38	6.74E-15	-0.50	4.00E-02	1.27E-12
rs6469265	RNA transport	84	28	2.46E-08	-0.46	2.70E-01	1.63E-05
rs6572658	Cell cycle	80	32	8.67E-11	-3.23	1.63E-01	4.78E-08
rs6683015	Protein processing in endoplasmic reticulum	105	21	2.26E-09	0.03	9.27E-01	5.59E-06
rs6687042	Cell cycle	80	27	3.36E-08	-4.66	7.10E-02	3.11E-06
rs6687042	RNA transport	84	30	1.25E-09	0.28	7.08E-01	2.41E-06
rs675679	Huntington's disease	120	50	2.56E-08	1.10	1.38E-01	9.85E-06
rs6964421	Melanogenesis	67	7	6.76E-04	-6.16	5.00E-06	6.24E-06
rs6967487	Protein processing in endoplasmic reticulum	105	37	2.79E-09	0.08	9.30E-01	7.12E-06
rs7014589	Cell cycle	80	28	1.87E-11	-2.56	4.19E-01	2.64E-08
rs7015262	Cell cycle	80	21	1.50E-10	-1.32	4.19E-01	1.71E-07
rs7093644	Huntington's disease	120	61	1.36E-09	0.59	2.08E-01	8.84E-07
rs7116631	Alzheimer's disease	101	48	2.56E-09	0.65	5.14E-01	9.26E-07
rs7116631	Cell cycle	80	41	2.21E-09	2.79	1.85E-01	4.04E-07

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs7116631	Huntington's disease	120	55	9.78E-10	0.59	2.50E-01	3.71E-07
rs7116631	Parkinson's disease	75	39	3.16E-09	0.00	9.96E-01	1.70E-06
rs7116631	Protein processing in endoplasmic reticulum	105	48	1.26E-08	0.56	5.30E-01	2.88E-06
rs7116631	RNA transport	84	44	2.28E-10	-0.39	1.91E-01	1.42E-07
rs7209818	Alzheimer's disease	101	31	9.03E-09	-1.28	2.25E-01	2.71E-06
rs7209818	Parkinson's disease	75	29	5.49E-11	-1.55	3.20E-01	5.75E-08
rs725229	Cell cycle	80	60	1.19E-12	-2.57	5.27E-01	1.25E-09
rs725229	RNA transport	84	65	9.82E-15	0.70	1.62E-01	7.64E-12
rs7260668	Huntington's disease	120	41	1.78E-09	-0.05	6.91E-01	1.70E-06
rs7260668	Parkinson's disease	75	35	9.64E-13	0.67	5.75E-01	2.09E-09
rs7351086	Protein processing in endoplasmic reticulum	105	23	1.50E-16	0.07	8.64E-01	3.99E-13
rs735738	RNA transport	84	34	2.71E-13	-0.45	1.04E-01	1.19E-10
rs7422930	RNA transport	84	51	7.01E-18	-0.30	3.42E-01	1.33E-14
rs752239	Cell cycle	80	36	3.34E-08	1.81	3.79E-01	1.63E-05
rs752239	RNA transport	84	41	1.71E-10	-0.42	1.19E-01	6.99E-08
rs7539844	Alzheimer's disease	101	23	1.12E-08	-1.39	1.81E-01	1.56E-06
rs7539844	Huntington's disease	120	31	8.69E-13	-0.40	7.31E-01	2.01E-09
rs7539844	Parkinson's disease	75	24	2.22E-12	-0.90	5.81E-01	2.01E-09
rs7586918	Protein processing in endoplasmic reticulum	105	16	4.56E-09	0.02	9.09E-01	7.66E-06
rs7616874	Protein processing in endoplasmic reticulum	105	28	7.43E-11	0.21	7.28E-01	1.65E-07
rs7621332	Huntington's disease	120	49	1.46E-09	2.34	3.69E-01	7.83E-07

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs7621332	Parkinson's disease	75	34	2.11E-08	0.00	9.99E-01	1.71E-05
rs7621332	RNA transport	84	40	1.91E-10	-0.64	5.04E-01	3.01E-07
rs7632176	RNA transport	84	53	3.75E-17	-0.30	3.45E-01	7.07E-14
rs7651414	Protein processing in endoplasmic reticulum	105	29	5.24E-13	0.39	5.45E-01	8.79E-10
rs765256	Alzheimer's disease	101	33	2.04E-09	0.59	4.23E-01	1.18E-06
rs765256	Huntington's disease	120	36	5.19E-09	0.20	7.57E-01	3.33E-06
rs765256	Parkinson's disease	75	29	2.09E-10	-1.06	2.82E-01	1.81E-07
rs7675985	Cell cycle	80	44	1.02E-09	-0.57	7.76E-01	1.18E-06
rs7675985	RNA transport	84	56	3.72E-17	0.30	3.99E-01	8.03E-14
rs7681425	Alzheimer's disease	101	29	3.10E-08	0.69	3.68E-01	9.16E-06
rs7681425	Huntington's disease	120	33	1.15E-08	0.20	8.91E-01	9.16E-06
rs7681425	Parkinson's disease	75	29	1.02E-11	0.19	8.29E-01	2.81E-08
rs7739002	Alzheimer's disease	101	32	1.14E-09	0.70	6.26E-01	1.04E-06
rs7739002	Parkinson's disease	75	29	2.95E-11	1.07	5.60E-01	5.63E-08
rs7780322	RNA transport	84	32	4.17E-09	-0.66	9.50E-02	1.15E-06
rs7794040	RNA transport	84	27	5.59E-09	-0.29	6.03E-01	9.06E-06
rs7844633	Alzheimer's disease	101	26	1.67E-10	0.58	3.42E-01	1.62E-07
rs7844633	Parkinson's disease	75	21	1.84E-09	0.80	3.40E-01	7.98E-07
rs7867279	Alzheimer's disease	101	46	7.57E-08	-1.07	1.75E-01	8.78E-06
rs7867279	Epstein-Barr virus infection	128	54	1.40E-07	-1.51	9.60E-02	8.78E-06
rs7867279	Huntington's disease	120	62	4.49E-13	-0.51	2.26E-01	4.30E-10
rs7867279	Parkinson's disease	75	42	9.47E-11	-0.41	6.92E-01	1.10E-07
rs7972875	Huntington's disease	120	44	4.16E-10	-0.28	4.90E-01	6.32E-07

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs8077875	Parkinson's disease	75	27	9.01E-10	2.05	4.48E-01	1.22E-06
rs8107491	Alzheimer's disease	101	24	5.96E-10	-1.36	3.60E-02	3.43E-08
rs8107491	Huntington's disease	120	26	9.90E-10	0.00	1.00E+00	8.96E-07
rs8107491	Parkinson's disease	75	25	6.32E-14	-0.85	3.80E-01	9.71E-11
rs8436	Alzheimer's disease	101	54	2.77E-10	-1.59	1.02E-01	2.43E-08
rs8436	Huntington's disease	120	69	7.33E-15	-0.39	3.61E-01	4.14E-12
rs8436	Parkinson's disease	75	53	3.10E-17	1.18	2.55E-01	4.34E-14
rs8436	RNA transport	84	55	1.56E-15	0.61	2.40E-02	9.85E-14
rs9263966	Huntington's disease	120	37	1.11E-10	-0.36	8.14E-01	2.88E-07
rs9263966	Parkinson's disease	75	27	7.93E-10	1.27	4.60E-01	5.47E-07
rs9299013	Huntington's disease	120	51	1.09E-08	-0.82	4.70E-02	1.57E-06
rs9374118	Protein processing in endoplasmic reticulum	105	25	6.97E-09	0.42	5.21E-01	9.20E-06
rs9398120	Alzheimer's disease	101	36	6.37E-08	-1.32	1.68E-01	6.78E-06
rs9398120	Huntington's disease	120	51	4.28E-14	-0.33	4.60E-01	4.19E-11
rs9398120	Parkinson's disease	75	43	4.23E-18	0.94	4.98E-01	1.15E-14
rs9398120	RNA transport	84	32	5.66E-08	0.61	4.50E-02	2.31E-06
rs9601213	Alzheimer's disease	101	28	1.31E-14	-0.37	5.52E-01	2.25E-11
rs9601213	Huntington's disease	120	26	5.46E-11	-0.19	8.67E-01	3.64E-08
rs9601213	Parkinson's disease	75	24	3.05E-14	-0.22	7.49E-01	3.44E-11
rs986475	Alzheimer's disease	101	37	2.04E-10	-1.47	2.13E-01	4.57E-08
rs986475	Huntington's disease	120	48	8.05E-15	-0.47	4.41E-01	7.73E-12

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pGFdr
rs986475	Parkinson's disease	75	39	5.45E-17	0.33	8.83E-01	2.36E-13
rs9896436	Cell cycle	80	24	4.16E-09	2.26	1.09E-01	1.21E-06
rs9901660	Huntington's disease	120	27	2.86E-08	-0.75	2.00E-02	6.87E-07
rs9901660	Parkinson's disease	75	23	4.92E-10	-1.06	4.43E-01	5.47E-07

APPENDIX B

COMPLETE REPLICATION ASSOCIATION RESULTS

This table contains all of the significant replication SNP-Pathway association results. For each SNP-pathway combination we report the following measures:

- pSize – The number of genes in the pathway with gene expression values available
- NDE – The number of differentially expressed genes in the pathway
- pNDE – The probability of observing the number of differentially expressed genes by chance alone.
- tA – The total accumulation of effect from differentially expressed genes in the pathway
- pPERT – The probability of observing the total accumulation value by chance alone
- pG – The combined probability of pNDE and pPERT.

Results are provided in order of increasing global p-value. SNPs with multiple pathway associations are not grouped together.

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pG
rs6572658	Cell cycle	119	56	7.14E-12	-6.79	0.14	2.84E-11
rs7681425	Parkinson's disease	105	53	2.41E-09	-4.08	0.02	9.07E-10
rs7681425	Huntington's disease	163	72	7.09E-09	-0.52	0.56	8.12E-08
rs11008749	Cell cycle	119	46	2.25E-08	4.43	0.32	1.41E-07
rs7972875	Huntington's disease	163	42	3.19E-08	-0.24	0.72	4.27E-07
rs12475079	Huntington's disease	163	40	3.37E-08	-0.38	0.75	4.64E-07
rs7867279	Epstein-Barr virus infection	177	52	0.03	5.54	5.00E-06	2.58E-06
rs10131614	RNA transport	126	41	4.62E-07	-0.04	0.82	6.01E-06
rs7586918	Protein processing in endoplasmic reticulum	150	42	3.23E-06	0.57	0.30	1.44E-05
rs425437	Huntington's disease	163	24	2.53E-06	0.00	1.00	3.51E-05

SNP	Pathway Name	pSize	NDE	pNDE	tA	pPERT	pG
rs7681425	Alzheimer's disease	148	60	4.04E-06	-0.03	0.99	5.37E-05
rs10517012	Olfactory transduction	361	17	1.00	6.66	5.00E-06	6.60E-05
rs1162371	Olfactory transduction	361	10	1.00	5.52	5.00E-06	6.60E-05
rs11104775	Olfactory transduction	361	22	1.00	12.07	5.00E-06	6.60E-05
rs11104947	Olfactory transduction	361	17	1.00	10.72	5.00E-06	6.60E-05
rs2688590	RNA transport	126	38	4.39E-05	-0.37	0.48	2.48E-04
rs6967487	Protein processing in endoplasmic reticulum	150	27	1.54E-04	0.36	0.53	8.48E-04
rs8436	Alzheimer's disease	148	18	1.52E-03	1.64	0.06	9.41E-04
rs735738	RNA transport	126	34	9.41E-04	-0.32	0.16	1.46E-03
rs6687042	Cell cycle	119	23	1.66E-03	4.52	0.17	2.63E-03
rs425437	Parkinson's disease	105	14	8.46E-04	0.00	1.00	6.83E-03
rs7780322	RNA transport	126	18	1.14E-03	0.00	1.00	8.87E-03
rs2526478	Cell cycle	119	25	0.57	3.21	3.00E-03	0.01
rs3750131	RNA transport	126	29	4.40E-03	-0.16	0.52	0.02
rs6687042	RNA transport	126	22	7.39E-03	-0.44	0.36	0.02
rs425437	Alzheimer's disease	148	16	3.51E-03	0.08	0.88	0.02
rs1388970	Cell cycle	119	8	4.33E-01	1.58	0.01	0.03
rs4626725	RNA transport	126	23	6.46E-03	-0.43	0.77	0.03
rs1634761	RNA transport	126	33	0.01	-0.10	0.49	0.03
rs4626725	Cell cycle	119	21	0.01	2.36	0.46	0.04
rs329312	Protein processing in endoplasmic reticulum	150	18	0.29	-0.87	0.03	0.04
rs8436	Huntington's disease	163	16	0.02	-0.33	0.42	0.05

APPENDIX C

COMPLETE FUNCTIONAL ANNOTATION RESULTS

This table contains all of the functional annotation results for each of the SNP-Gene-Pathway associations. This table contains a variety of information:

- DNaseI sensitivity: Cell line types for which the SNP is found in open chromatin.
- Chromatin State: Interpreted chromatin function for the given cell type/s.
- Transcription Factor Binding Sites (TFBS): The transcription factor and the cell line tested.
- Altered Regulatory Motif: The relative affinity level of the given transcription factor for the minor allele compared to the reference allele.
- RegulomeDB score: Level of evidence supporting regulatory function.
    - 1f: evidence for an eQTL and either transcription factor binding or a DNase hypersensitivity peak.
    - 5: Either transcription factor binding or a DNase hypersensitivity peak at the variant.
    - 6: "Other" category indicating minimal binding evidence

When an annotation type is not available, the entry is shaded with a grey diagonal pattern.

SNP / Gene Pathway	DNaseI Sensitivity	Chromatin State	TFBS	Altered Regulatory Motif	Regulome DB Score
rs425437 / MOSC2, C1orf115 Huntington's Disease	• LCL • Glioblastoma • Prostate adenocarcinoma • Primary tracheal epithelial cells	• Enhancer — Brain Inferior Temporal Lobe — Brain Angular Gyrus • Weak Enhancer — H1 Derived Mesenchymal Stem Cells — Pancreas — Spleen • Transcription Enhancer-like — Gastric • Transcription Enhancer-like (short gene) — H1 BMP4 Derived Trophoblast Cultured Cells		• Increased affinity: — HIF1	1f
rs7586918 / DTNB Protein processing in ER	• LCL • Primary Th1 and Th2 T cells • Chronic lymphocytic leukemia • Medulloblastoma • Osteoblasts • Urothelial cells • CD4+ cells • B cells • hematopoietic progenitor cells	• Strong Enhancer — LCL • Active Enhancer — Mobilized CD34 Primary Cells • Enhancer — Spleen • Weak Enhancer — Mobilized CD34 Primary Cells — CD34 Primary Cells — CD3 Primary Cells — CD8 Naïve & Memory Primary Cells — CD19 Primary Cells	• LCL — PBX3 — POL2 — POL24H8 — PU1 — SIN3AK20 — SRF — TBP — TCF12		1f

SNP / Gene Pathway	DNaseI Sensitivity	Chromatin State	TFBS	Altered Regulatory Motif	Regulome DB Score
rs7867279 / ZNF510 Epstein-Barr virus inf.	• Retinal pigment epithelial cells • Fibroblasts, Hutchinson-Gilford progeria syndrome	• Weak Enhancer – Fetal Heart		• Reduced affinity: – NF-1	5
rs6572658 / L2HGDH Cell cycle		• Weak Enhancer – Fetal Heart		• Reduced affinity: – Evi-1_2 – HDAC2_disc6	No Data
rs7681425 / STIM2 Hunt., Park., Alz. Disease	• Prostate adeno-carcinoma	• Weak Enhancer – Adult Liver			5
rs10131614 / EIF2S1 RNA transport				• Reduced affinity: – COMP1 – FAC1 – Foxa – Foxd3 – Foxk1 – Foxo_2 – Foxp1 – Nanog – Sin3AK-20_disc3 – Sox_13 – Sox_2 – Sox_6 – ZFP105 – P300_disc5 • Increased affinity: – HMG-IY_2 – RREB-1_2	6

SNP / Gene Pathway	DNaseI Sensitivity	Chromatin State	TFBS	Altered Regulatory Motif	Regulome DB Score
rs11008749 / ARHGAP12 Cell cycle	• Epidermal keratinocytes	• Active Enhancer — CD4+ CD25- IL17+ PMA-Ionomcyin stimulated Th17 Primary Cells • Weak Enhancer — CD4+ CD25- CD45RA+ Naive Primary Cells — CD8 Naive Primary Cells — CD4 Memory Primary Cells • Transcription Enhancer-like — CD4+ CD25- Th Primary Cells — CD4+ CD25- IL17- PMA-Ionomycin stimulated MACS purified Th Primary Cells — CD4+ CD25- CD45RO+ Memory Primary Cells — Mesenchymal Stem Cell Derived Adipocyte Cultured Cells — CD15 Primary Cells — CD4 Naive Primary Cells — Colon Smooth Muscle — Penis Foreskin Melanocyte — CD8 Memory Primary Cells — CD3 Primary Cells		• Increased affinity: — OCT-1 — ZFP187	1f

SNP / Gene Pathway	DNaseI Sensitivity	Chromatin State	TFBS	Altered Regulatory Motif	Regulome DB Score
rs12475079 / IMMT Huntington's disease	• Choroid plexus epithelial cells • LCL • epithelial cell line from lung carcinoma • embryonic stem cells • hepatocellular carc. • leukemia • mammary gland, adenocarcinoma • epid. keratinocytes • fetal buttock/thigh fibroblast • gingival fibroblasts • promyelocytic leukemia cells • blood microvascular endothelial cells, lung • neonatal blood microvascular endothelial cells, dermal • renal cortical epithelial cells • T lymphoblastoid • acute promyelocytic leukemia • malignant pluripotent embryonal carcinoma • renal proximal tubule epithelial cells • primary Th2 T cells	• Active Promoter – HepG2 • Weak Promoter – LCL; NHLF; HMEC; Huvec; NHEK; HSMM; H1 • Strong Enhancer – K562 • Active Enhancer – CC.TPC; R.MUC31; CCCRA.NP; CCC.TREGP; CCIP.LSTP; CD34.MBP1536; LNG.FE; BN.FE2; BN.AC; SPL; ST.MUC; CD19.P; CCCRO.MP • Enhancer – ESO; CCC.TMP; H1.DMSC • Weak Enhancer – CD4.NP; CD8.NP; CD34.MBP1480; H1.BMP4DM; PANC; H1; H9 • Transcription Enhancer-like – GAS • Transcription Enhancer-like (short gene) – PFK.3; HD.CD56MESC • TSS-Flanking more upstream – PFK.2; PFF.1; ADI.MSC; MSC.ADIPC;	• CTCF – BJ – Caco-2 – LCL – HBMEC – HCPEpiC – HSMM tube – HepG2 – K562 – MCF-7 – NHDF-Ad – NHEK – Osteobl	• Reduced affinity: – LUN1	1f

SNP / Gene Pathway	DNaseI Sensitivity	Chromatin State	TFBS	Altered Regulatory Motif	Regulome DB Score
		HRT.FE; NCC.GED2; CD34.MBP1508; CCIP.LSMPTP; MUS.SC; CD3.P; CD34.MBP1562; CD34.C; PFM.1,2,3; BR.H35; BN.GM2; ES.I3; LV; BM.MSC; CHON.BMMSC; PAN.I; CD4.MP; R.MUC29; SK.MUS63; CD15.P; NCC.COR2; ADI.NUC; CD8.MP; DUO.SMUS • TSS Active – PFF.2 • TSS Weak – HD.CD184EC • TSS Flanking more downstream – IMR90; R.SMUS; SK.MUS; BN-FEO; CD34.P; BN.FE1; IPS.20; ST.SMUS28; COL.SMUS; KID.FE; BN.HM150; BN.AG DUO.MUC61; BN.CC; BN.ITL; BN.MFL; BN.SN; HUES48; HUES64; HUES6, IPS.15; CD34.MBP1549; LIV.A; BR.MYO; COL.MUC32			

SNP / Gene Pathway	DNaseI Sensitivity	Chromatin State	TFBS	Altered Regulatory Motif	Regulome DB Score
rs7972875 / VPS33A Huntington's disease	• LCL			• Reduced affinity: – NF-1	5
rs1162371 / CEP290 Olfactory transduction		• Enhancer – BN.GM2 • Weak Enhancer – NCC.COR2; NCC.GED2; BN.MFL; BN.FE1		• Reduced affinity: – STAT3	6
rs11104775 / CEP290 Olfactory transduction		• Enhancer – MSC.ADIPCl; PFF.2 • Weak Enhancer – Huvec; NHLF; ADI.MSC; PFF.1		• Reduced affinity: – CDP_4 – Dlx2 • Increased affinity: – Bsx	6
rs11104947 / CEP290 Olfactory transduction		• Active Enhancer – PFF.1 • Weak Enhancer – PFF.2; ADI.MSC			6
rs10517012 / TMEM33 Olfactory transduction		• Poised Enhancer – PFF.1; MSC.ADIPC • Weak Enhancer – LIV.A; ADI.MSC		• Reduced affinity: – HNF4_disc4	6